Date of acceptance Grade

Instructor

Deep Groundwater Metagenomics - Computational Analysis of Microbial Communities and Metabolic Pathways

Nicole Althermeler

Helsinki April 3, 2015

M.Sc. Thesis

UNIVERSITY OF HELSINKI
Department of Computer Science

HELSINGIN YLIOPISTO — HELSINGFORS UNIVERSITET — UNIVERSITY OF HELSINKI

Tiedekunta — Fakultet — Faculty	Laitos — Institution — Department
Faculty of Science	Department of Computer Science

Tekijä — Författare — Author
Nicole Althermeler

Työn nimi — Arbetets titel — Title

Deep Groundwater Metagenomics - Computational Analysis of Microbial Communities and Metabolic Pathways

Oppiaine — Läroämne — Subject
Bioinformatics

Työn laji — Arbetets art — Level	Aika — Datum — Month and year	Sivumäärä — Sidoantal — Number of pages
M.Sc. Thesis	April 3, 2015	65 pages

Tiivistelmä — Referat — Abstract

Metagenomics promises to shed light on the functioning of microbial communities and their surrounding ecosystem. In metagenomic studies the genomic sequences of a collection of microorganisms are directly extracted from a specific environment. Up to 99% of microbes cannot be cultivated in the lab; thus, traditional analysis techniques have very limited applicability in this challenging setting. By directly extracting the sequences from the environment, metagenomic studies circumvents this dilemma. Thus, metagenomics has become a powerful tool in the analysis of the diversity and metabolic capability of environmental microbes. However, metagenomic studies have challenges of their own.

In this thesis we investigate several aspects of metagenomic data set analysis, focusing on means of (1) verifying adequacy of taxonomic unit and enzyme representation and annotation in the sample, (2) highlighting similarities between samples by principal component analysis, (3) visualizing metabolic pathways with manually drawn metabolic maps from the Kyoto Encyclopedia of Genes and Genomes, and (4) estimating taxonomic distributions of pathways with a novel strategy.

A case study of deep bedrock groundwater metagenomic samples will illustrate these methods. Water samples from boreholes, up to 2500 meter deep, of two different sites of Finland display the applicability and limitations of aforementioned methods. In addition publicly available metagenomic and genomic samples serve as baseline references.

Our analysis resulted in a taxonomic and metabolic characterization of the samples. We were able to adequately retrieve and annotate the metabolic content based on the deep bedrock samples. The visualization provided a tool for further investigation. The microbial community distribution could be characterized on higher levels of abstraction. Previously suspected similarities to fungi or archaea were not verified. First promising results were observed with the novel strategy in estimating taxonomic distributions of pathways.

Further results can be found at: http://www.cs.helsinki.fi/group/urenzyme/deepfun/

ACM Computing Classification System (CCS):
Applied computing ∼ Bioinformatics
Applied computing ∼ Computational biology
Applied computing ∼ Genomics
Applied computing ∼ Metabolomics / metabonomics

Avainsanat — Nyckelord — Keywords
Bioinformatics, Metagenomics, Computational Biology

Säilytyspaikka — Förvaringsställe — Where deposited

Muita tietoja — övriga uppgifter — Additional information

Acknowledgment

I would like to express my gratitude to my supervisor Professor Juho Rousu for his continuous support and patience through the learning process of this master thesis. I would also like to thank my supervisor at the University of Helsinki Professor Mikko Koivisto for or helpful comments in finalizing this manuscript. Furthermore, I would like to thank Hongyu Su for his instructions and my offices mates and colleagues at Aalto University for the great lunches.

I wish to thank the project collaborators at VTT, The Technical Research Center of Finland. My gratitude goes to Merja Itävaara, and a special thanks goes to Fahad Syed, for his work on the assembly of the sequencing data, and to Heikki Salavirta, for his work on preprocessing the data.

Abschließend möchte ich noch meiner Familie für ihre immerwährende Unterstützung danken.

Contents

1 Introduction and Background

Microbial communities exist in all ecosystems, such as groundwater and soil as well as the skin surface and the gastrointestinal tract. It is estimated that our planet is inhabited by 5×10^{30} prokaryotic cells [WCW98]. Alone the human body holds more bacterial cells (10^{14}) than it itself consist of (10^{13}) [Sav77, Ber96].

The microbes of the communities interact with each other and take part in the biological processes surrounding them. A profound expertise of the habitat can help estimating effects changing or new outside influences have on the complex system of the microbial communities.

Microbial communities can be characterized by various geochemical and biochemical measurements. One possible means of investigation is sequencing the genomic material with next generation sequencing (NGS). In the analysis of individual microbes, increased throughput and lower cost of NGS technologies promoted scientific achievements with great impact [KSL+13]. The availability of NGS has lead to a rapid increase of generated data, challenging data storage, management and analysis. Various computational methods are developed to analyze NGS data, in the early stages for genomic and nowadays also for metagenomic data. Pipelines are generated to gather such analysis tools and easily apply a standard procedure of common analysis techniques.

In recent years approaches integrating data from several experiments have been used to consider complex mechanism from different angles. Epigenomics, transcriptomics, proteomics and genomics each serve as a different powerful angle, which are combined capable of unraveling complex processes and organisms [HHR10].

The goal of this thesis is to characterize the deep groundwater microbial communities by means of metagenomic analysis. Several established as well as novel techniques will be applied to preprocessed data sets. These methods focus on means of (1) verifying the adequacy of taxonomic unit and enzyme representation in the sample, (2) highlighting similarities between samples by principal component analysis, (3) visualizing metabolic pathways with manually drawn metabolic maps from the Kyoto Encyclopedia of Genes and Genomes, and (4) estimating taxonomic distribution of pathways.

We will analyze metagenomic sequencing data from deep groundwater samples which were taken from two different sites and different depths in Finland, Olkiluoto and Outokumpu. The sample from Olkiluoto was taken from groundwater fractures and

the samples from Outokumpu where taken from different depths of the same deep bore hole. The samples were taken directly from their environment. Therefore they can be compared in respect of the microbes and the living conditions surrounding them.

In the following background knowledge crucial for the course of this thesis is presented.

1.1 Metagenomics

The term metagenomics was first coined in 1998 [HRB+98] and is nowadays defined as the direct extraction of the whole genome information from a habitat and its following analysis. The microbial community of a habitat typically consists of microorganism, such as bacteria, some protozoans, archaea and fungi, as well as bacteriophages.their own metabolism, are often not classified as microorganism.

Before the advent of metagenomics, samples of one habitat were analyzed by first cultivating the individual microorganisms separately. However, as seen in the "Great count anomaly"[SK85], it is believed that up to 99% of microorganism in environmental samples cannot be cultured with currently available technologies [ALS95]. In the past, this lead to a limited and biased view on the microbial communities which was improved with the emerge of metagenomic studies. For instance, one of the first metagenomic study greatly improved our view on naturally occurring marine plankton by discovering bacterial rhodopsins that function as light-driven proton pumps [BAK+00]. Later, the diversity of microorganism in water environments was revealed by two major studies [VRH+04, RHS+07], which identified 148 novel phylotypes and ∼360 species as well as predicted ∼7 million novel genes.

Metagenomic analysis entails several steps. The first step is analyzing the extraction and sequencing of DNA. Sequencing approaches can be divided into whole (meta)genome shotgun sequencing, which randomly sequences the mixed genomes, and target sequencing, which only sequences specifically targeted gene(s).Typical targets are the 16S rRNA gene or genomic regions such as the Internal Transcribed Spacer (ITS) regions, as they allows for species level classification. Targeted sequencing is more cost and time efficient, as well as more specific and exact, but it also assumes some existing knowledge of the sample at hand and the results are useful only for a limited number of applications.

Given the sequencing reads of a whole genome shotgun sequenced sample, two dif-

ferent strategies of genome reconstruction exist. A first strategy is to assemble the reads into longer, more specific contigs and use these less error-prone sequences as the basis for taxonomic classification and functional analysis. A second strategy works on the initial short read which avoids chimeric assembly, but leads to a high error rate and a big quantity of reads to be processed.

Various projects focus on obtaining open-source metagenomic samples. For instance, *The Earth Metagenomic Project* http://www.earthmicrobiome.org/ focuses on soil samples and *The Human Microbiome Project*, http://commonfund.nih.gov/hmp/index, focuses on disease related body sites, namely oral cavity, nasal cavity, skin, gastrointestinal tract and urogenital tract.

Metagenomics still faces several problems. The sample preparation and the sequencing method can cause biased results. The read lengths generated during sequencing influence sequence assembly, gene prediction and subsequently further analysis. In addition, analysis results are potentially affected by the sample inherent characteristics, such as the average genome size of the organism of the sample. preparation, [TG12, PT12] .

1.1.1 Preparation of Metagenomic Samples

Assembly of metagenomes. The de novo assembly of metagenomes is an important but challenging step in metagenomic analysis [PS10]. During the assembly shorter, overlapping reads are combined to larger contigs. In the past, single genome studies reconstructed one genome at a time and the main goal was a complete gap less sequence. However, nowadays metagenomic studies try to reconstruct the genomes of multiple organisms at the same time. Some of the organisms are closely related and have very similar genomes. The similarities in their genomes unavoidably lead to chimeras, contigs based on reads originally from multiple organisms, which may not even classify in the same phyla [MIB$^+$07]. In addition, the probability of finding overlapping reads is low in many environments and strongly depends on how well the genomes are coverage by the sequenced reads. [SH05].

Removal of Contamination. Contamination with human DNA or other material may occur during the sampling and DNA extracting process of genomic analysis. Therefore, screening the initial contig or read data sets is a necessary step. There exist a few tools designed for removal of contamination from metagenomics data sets, such as DeconSeq [SE11]. Based on the sample and its preparation process,

different cleaning steps may have to be applied.

1.2 Phylogeny and Taxonomy

Phylogeny is the evolutionary history of a species or groups of related species. It is represented in the 'Tree of Life'; a hierarchical structure in which every life-form is represented in relation to every other life-form. A first drawing of this concept was created by Darwin, a reproduction of one of the first 'Tree of Life' representations by is shown in Figure 1.

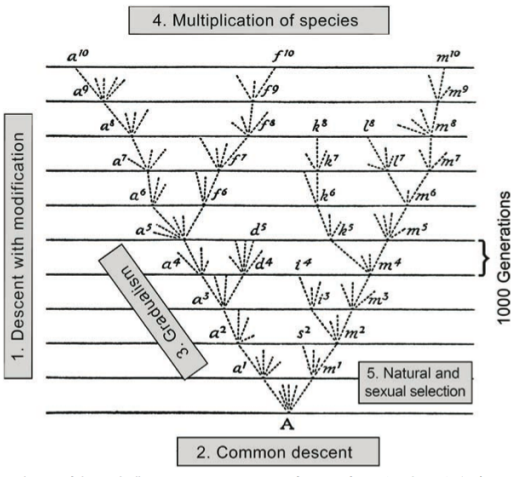


Figure 1: Partial reproduction of an illustration in Darwin's Origin of Species of 1859 (6. ed. 1872). Taken from Kutschera's ''From the scala naturae to the symbiogenetic and dynamic tree of life" [Kut11]. Five Darwinian species theories are added to illustrate the evolution of species from a common ancestor (A).

Taxonomy is the identification, naming and classification of the diverse forms of life. Classifications can be done with different levels, or ranks of specificity, with species being the most specific. At any given level, a named taxonomic unit is called a taxon [RUC+10]. Initially, Carl von Linnaeus, the father of modern taxonomy as we know it, proposed a simple five-level system: kingdom > class > order > genus > species [vLS59]; however, many additional intermediate levels were added later. The most notable intermediate levels are the domain (placed above the kingdom), the phylum (ordered between kingdom and class), and the family (between order and genus).

Taxonomy of metagenomic samples. The microbiomes in metagenomic samples can be highly variable in their taxonomic composition [TCH+04]. The approach for the characterization of taxonomic diversity is designated by the basic concept of the used sequencing method. For the determination of the taxons, the sequencing method of choices is targeted sequencing. Commonly ribosomal rRNA is used as a marker gene for taxon identification. The other approach, based on whole genome shotgun sequencing, uses short, random reads as representations of the composition of their source genomes.

Regardless of the sequencing method, the generated sequence reads, either targeted or random, are grouped into bins. This process, called binning, is either taxonomy dependent or taxonomy independent [MMG12]. Taxonomy dependent methods assign read annotation based on similarity to sequences/models of known phylogenetic origin. These methods highly depend on available reference genomes in databases as well as obtaining sufficient similarity. Thus, many reads might stay unassigned. All taxonomy dependent binning methods should therefore only be considered as estimations of the microbiome compositions. Methods which are independent of taxonomy, on the other hand, bin reads based on mutual similarities and don't require reference databases [MMG12].

Several classification and binning tools exist. Taxonomic depend binning tools are MG-Rast [GWW+10], MEGAN[HMW+11], PhyloPythiaS[PHP+11] and PhymmBL [BS09], to name a few. Fewer taxonomic independent binning tools exist; some representatives are TETRA[TWL+04] and MetaCluster[LYY+11].

1.3 Metabolism

Life relies on thousands of biochemical reactions that occur within a cell. The the complete set of chemical reactions that manage the material and energy resources of an organism is referred to as metabolism. The vast number of biochemical reactions are almost all connected to each other, the final products of one or multiple reaction function as the starting substances of the next reaction, thus creating a complex network. A series of chemical reactions that either build or break down a complex molecule is called a metabolic pathway [RUC+11, AJL+07].

Metabolic reactions can be divided into two main categories: Anabolism and catabolisms. Anabolism connects simple molecules to more complex ones, for instance the synthesis of proteins from amino acids is an anabolic reaction. The required energy is

stored in the new compound molecule. Catabolism is the complement of anabolism. Complex molecules are degraded and energy is released. For instance, the polysaccarid starch as found in many food products, is hydrolyzed in order to gain energy [RUC+11, AJL+07].

Enzymes play a key role in the monitoring of metabolism. The Chemical reactions in the cell would normally take place at a much higher temperature, but enzymes catalyze, that is to say increase the rate, of reactions taking place. Enzymes are specialized proteins, thus can be linked back to genes in the genome, meaning their quantitative and temporal occurrence is controlled by the cell. There are, however, a few exceptions: Some chemical reactions are non enzymatic and occur spontaneously [CF11].

Metabolism of metagenomic samples. Many pipelines offer solutions for functional annotation of metagenomic samples, such as CAMERA [SCL+11], MG-RAST [GWW+10], IMG-M [MCC+12]. The most straightforward and most common approach for functional annotation, which is implemented in most available pipelines, is the homology-based approach. This approach depends on a comprehensive database of already well annotated genes.

Another approach for functional annotation is motif or pattern based, in which the focus is moved to structural similarity instead of, despite similar function, sometimes diverse sequences. Databases such as PROSITE [SCdC+10] and PRINTS [ABF+03] and pipelines such as IMG-M [MCC+12] support this approach. Furthermore, it is possible to apply a context-based approach, for instance genomic neighborhood [DSHB98], or an approach inferring the putative role of a protein, such as membrane proteins with HMM-TM [BLH06] or lipoproteins with LIPO [BKS+06].

Typically, homology-based approaches, that are performed against publicly available reference sequence datasets, result in very reliable functional annotations for metagenomic sequences [PT12]. However, it should be noted that this approach suffers from the short read length of current next generation sequencing (NGS). The average read length of Illumina is $\sim$ 300bp [Ill14], significantly shorter than the average size of $\sim$ 1000bp of a protein [MMT+07].

2 Data sets and Preparation

In this chapter we will introduce all data sets used throughout the thesis. The data sets stem form several sources and may either be metagenomic or individual genomic samples. The sequencing data underwent preparation steps as well as functional and taxonomic annotation, which is described in Section 2.2 of this Chapter.

2.1 Description of Data

2.1.1 Metagenomic Samples from Deep Boreholes in Finland

Six novel metagenomic water samples were taken from two different sites in Finland. The samples OLKR40 and OLKR49 were taken from groundwater fractures at ∼600m and ∼500m in Olkilouto, Finland and sequenced by 454 total sequencing technique. The samples OUTO3, OUTO4, OUTO5, OUTO6 were taken from 967m, 2260m, 500m, 2300m, respectively, in a deep borehole located in Outokumpu, Finland. DNA extraction for the OUTO samples was done using two different DNA sample preparation kits, Trueseq (OUTO3 and OUTO5) and Nextera (OUTO4 and OUTO6). All OUTO samples were sequenced by the Illumina paired-end sequencing method with the Illumina GenomeAnalyzer. An overview of these data sets can be found in Table 1.

2.1.2 Publicly Available Metagenomic Samples

Two additional water metagenomic samples were used as baseline references. The *GW* sample metagenomic data set is from groundwater microbial community from a contaminated well in Oak Ridge, Tennessee [HDG+10]. The sequence reads were assembled into 421 contigs. The MAAOC sample was retrieved from a freshwater propionate anammox bacterial community from a bioreactor in Nijmegen, Netherlands [Ins14]. The samples were sequenced with the short insert Sanger and 454-Titanimum technique. An overview of these data sets can be found in Table 2.

2.1.3 Publicly Available Genomic Samples

In addition to metagenomes, genomic material from multiple individual microbes were also used as reference data sets. We selected a set of 39 publicly available

Sample	Origin	depth	Sequencing	DNA kit	Assembly
OUTO3	Outokumpu, Finland	967m	llumina paired-end	Truseq DNA kit	Yes
OUTO4	Outokumpu, Finland	2260m	llumina paired-end	Nextera DNA	Yes
OUTO5	Outokumpu, Finland	500m	llumina paired-end	Truseq DNA kit	Yes
OUTO6	Outokumpu, Finland	2300m	llumina paired-end	Nextera DNA	Yes
OLKR40	Olkiluoto, Finland	$\sim$ 600m	454		No
OLKR49	Olkiluoto, Finland	$\sim$ 500m	454		No

Table 1: Overview of metagenomic samples from deep boreholes in Finland

Sample	Origin	Sequening	Assembly
GW	groundwater microbial community, contaminated well Oak Ridge, Tennessee	-	Yes
MAAOC	freshwater propionate anammox bacterial community,bioreactor Nijmegen, Netherlands	short-insert Sanger and 454	-

Table 2: Reference metagenomic data sets

genomes covering the domains archaea, bacteria and the eukaryotic kingdom of fungi. From the domain bacteria we considered the following four phyla: Proteobacteria, Firmicutes, Chlorobacteria, and Bacteroidetes. The Proteobacteria can be further classified into Alpha, Beta, Gamma, Delta, and Epsilon class. The Bacteroidetes are represented only by the class of Flavobacteria.

An overview of the genomic data sets, as well as their unique assession number under which they have been submitted to DDBJ/EMBL/GenBank databases [BvdBC$^+$00], can be found in Table 3, 4, and 5.

Sample ID	Name	Class
NC 007205.1	Alpha Candidatus Pelagibacter ubique HTCC1062	Alpha
NC 010505.1	Alpha Methylobacterium radiotolerans JCM 2831	Alpha
NC 008752.1	Beta-Acidovorax citruli	Beta
NC 011992.1	Beta-Acidovorax ebreus TPSY	Beta
NC 014207.1	Beta-Methylotenera versatilis 301	Beta
NC 012968.1	Beta-Methylotenera mobilis JLW8	Beta
NC 016830.1	Gamma Pseudomonas fluorescens F113	Gamma
NC 009512.1	Gamma Pseudomonas putida F1	Gamma
NC 002977.6	Gamma Methylococcus capsulatus str. Bath	Gamma
NC 010943.1	Gamma Stenotrophomonas maltophilia K279a	Gamma
NC 014972.1	Delta Desulfobulbus propionicus DSM 2032	Delta
NC 014844.1	Delta Desulfovibrio aespoeensis Aspo-2	Delta
NC 017454.1	Delta Geobacter sulfurreducens KN400	Delta
NC 007575.1	Epsil Sulfurimonas denitrificans DSM 1251	Epsil
NC 014762.1	Epsil Sulfuricurvum kujiense DSM 16994	Epsil
NC 008554.1	Epsil Syntrophobacter fumaroxidans MPOB	Epsil

Table 3: Reference genomes of proteobacteria

Sample ID	Name	Phylum
NC 009441.1	Flavo Flavobacterium johnsoniae UW101	Flavo
NC 009613.1	Flavo Flavobacterium psychrophilum JIP02/86	Flavo
NC 014960.1	Chlor Anaerolinea thermophila UNI-1	Chlorobacteria
NC 010175.1	Chlor Chloroflexus aurantiacus J-10-fl	Chlorobacteria
NC 008346.1	Firmi Syntrophomonas wolfei	Firmicutes
NC 013520.1	Firmi Veillonella parvula DSM 2008	Firmicutes
NC 013216.1	Firmi Desulfotomaculum acetoxidans DSM 771	Firmicutes
NC 018017.1	Firmi Desulfitobacterium dehalogenans ATCC 51507	Firmicutes
NC 009253.1	Firmi Desulfotomaculum reducens	Firmicutes
NC 018068.1	Firmi Desulfosporosinus acidiphilus SJ4	Firmicutes

Table 4: Reference genomes of bacteria

2.2 Preparation of Data

The data sets introduced above are in a wide range of qualities. This section focuses on the processing steps the different samples underwent in order to be comparable.

Sample ID	Name	Domain
NC 000916.1	Archaea Methanothermobacter thermautotrophicus str. Delta H chromosome	Archaea
NC 003552.1	Archaea Methanosarcina acetivorans C2A chromosome	Archaea
NC 015416.1	Archaea Methanosaeta concilii GP6 chromosome	Archaea
NC 002689.2	Archaea Thermoplasma volcanium GSS1 chromosome	Archaea
CD 000001.1	Fungi cryptococcus neoformans grubii h99 2 contigs	Fungi
CD 000002.1	Fungi neurospora crassa or74a finished 10 contigs	Fungi
CD 000003.1	Fungi ustilago maydis 1 contigs	Fungi

Table 5: Reference genomes of archaea and fungi

The overall goal is the retrieval of the following matrices, which are the starting point and input for the methodologies presented in the main part of this thesis.

- A sample $\times$ enzyme matrix containing the number of times each enzyme was found in each sample.
- A sample $\times$ enzyme matrix containing the best corrected blast bit scores for each enzyme in each sample.
- For each sample: a enzyme $\times$ taxonomic unit matrix on each taxonomic level.

In the following, further details about the construction of these matrices are given.

2.2.1 Preparation of the Novel Metagenomic Samples from Finland

challenging step in metagenomic analysis [PS10]. In the past, single genome studies reconstructed one genome at a time and the main goal was a gapless sequence. Metagenomic studies on the the other hand tries to reconstruct the genomes of multiple organisms at the same time, some of which are closely related and thus show similarities in their genomes. The similarities in their genomes will unavoidable lead to chimeras, contigs based on reads originally from multiple organisms. These organisms may not even classify in the same phyla [MIB$^+$07]. In addition, the probability of finding overlapping reads is low in many environments and strongly depend on how well the genomes are coverage by the sequenced reads. [SH05].

Performed assembly. For the novel OLKR samples, the read distribution was too sparse to generate an assembly. The OUTO samples, on the other hand, covered the genomes sufficiently so that an assembly was possible. Here three different tools were used: MetaVelvet [NHTS12], MetaIDBA [PLYC11] and String Graph Assembler (SGA) [SD12]. Each of them has a different approach to assembly. Each tool was applied to our four samples and for each sample the best resulting assembly was chosen based on the best overall assembly size, the largest found contig and the best weighted median of the average length of a set of sequences, also called N50 value.

Performed removal of contamination. For this study, a work-flow specifically designed for the data was used. This workflow is based on homologies to the NCBI's protein database. All reads or contigs which align using BLAST to a human or eukaryotic associated sequence in the database, but not to a bacteria or archaea, were removed. A high identity of 95% and an alignment length of $\geq$ 100 bp prevent incorrect removal of contigs or reads.

2.2.2 Preparation of the Reference data sets

The publicly available (meta)genomes were sequenced in various laboratories and different methods for DNA preparation, sequencing and further processing have been applied to them. We assume that the reads of the publicly available data sets have been subject to the best possible processing and assembly to contigs.

2.2.3 Searching for Homologies in Annotated Protein Databases

BLAST [AGM$^+$90], the Basic Local Alignment Search Tool, is a widely used algorithm to find similar regions between nucleotide or protein sequences. Using this tool, the reads or contigs of the data sets were compared with the Uniprot protein knowledge base [JBD$^+$09]. Only sequences from the manually annotated and reviewed Swiss-Prot section of Uniprot were used in order to increase the reliability of the protein annotation. The proteins in the Uniprot database are represented by their amino acid sequence. However, our reads and contigs are nucleotide sequences. BLASTx is a special BLAST program which translates the nucleotide sequence into the corresponding 6 amino acid sequences and then compares it to a protein database. We used this tool for finding similar proteins to those encoded

by our nucleotide reads and contigs. For each read or contig, BLAST gives a list of hits of similar proteins and their alignments. Additionally the list includes, the origin species as well as statistical values like a percentage of identity, the matching length of the read, a bit score and an e-value. The bit score shows the validity of the alignment, where a higher score means a better alignment. The e-value is calculated based on the bit score and is the number of alignments expected by chance with the particular score or better. we only keep hits above certain, empirically determined, threshold in order to avoid using proteins for our further analysis with low similarity. All hits had to have a matching length of at least 30 base pair, a percentage of identity of at least 40% and an e-value below 1×10^{-6}.

2.2.4 Correcting of BLAST Bit Score

In this section we want to normalize and correct the BLAST bit scores to our further needs. Three main issues are considered:

1. Searching a metagenomic sample against a database is an alteration of the original purpose of BLAST, which is simply searching for a single read in a database. Now, we blast a set of reads, the metagenomic sample against a database, thus testing for multiple reads if they are in the database. By doing this, we increase the chance that we observe a hit as true, although it did just appear by chance. There is a need to correct the statistical values given by BLAST accordingly.

2. Some of the sequence reads may hit to the same protein. All hits will contribute to our belief that the protein in question is present in our sample. The reads do not overlap, or the overlap is not sufficiently strong to result in a contig during assembly, or the reads originate from different species. Nevertheless, we detected said protein in our sample multiple times and do want to give a bit score accordingly.

3. The proteins in the database differ in their sequence length. A complete alignment to a shorter protein will always result in a smaller bit score than a complete alignment to a longer protein. A shorter sequence is more likely to be generated by chance. For our purposes, we need values in the same range for each protein, rather focusing on the perceptual length of each protein being covered.

As stated in Section 2.2.3, each alignment is assigned a bit score and e-value. They help to determine the evidence of the homology and the likelihood of an alignment to have arisen by chance. The likelihood is based on the prospect that a certain sequence is generated randomly based upon a protein sequence model alone.

Both of these values, bit score and e-value, are based on the high scoring pairs (HSPs). HSPs are segment pairs which cannot be extended or trimmed anymore in order to achieve a higher score. The score for an HSPs, in the following denoted as S^{raw}, is calculated as

$$\lambda S^{raw} = \sum_{pos=1}^{l} \lambda S_{pos}^{raw}$$

where λS_{pos}^{raw} is the score for a single amino acid match, calculated as

$$\lambda S_{pos}^{raw} = log \frac{p_{ij}}{q_i q_j}$$

and where:

- l is the length of the HSP.
- p_{ij} is the amino acid frequency of a match between i and j taken from a scoring matrix such as BLOSUM62 [HH92].
- q_i and q_j are the frequency of amino acid i and j, respectively.
- λ is a scoring matrix dependent normalization factor.

The Blast bit score is calculated as:

$$S^{bit} = \frac{\lambda S^{raw} - ln(k)}{ln(2)}$$

where k is a further scoring matrix dependent constant.

The E-value E is then calculated as:

$$E = kmne^{\lambda S^{raw}}$$

where n and m are the effective length of the read sequence and database sequences.

BLAST bit score correction Under the current BLAST scoring scheme, the bit score is independent of the read length and the size of the target database. This is a disadvantage because bit scores from different searches are not comparable since the different read length are result in different search space. The E-value is dependent of the read length and the size, however, it is unsuitable to be corrected for the other issues mentioned above.

Search space correction for individual HSPs. (Issue 2) We adjust the bit score by considering the actual search space according to

$$S_{new}^{bit} = S^{bit} - \frac{ln(mn)}{ln(2)} \tag{1}$$

where n and m are the effective length of the read sequence and database sequences.

Combining of HSPs. (Issue 1) In many cases, BLAST aligns one read sequence to one protein with multiple ordered non-overlapping HSPs. If we combine the HSPs, we need to consider the gaps between them, as well. We calculate a bit score S_{cor}^{bit} of a read-protein pair from alignment score of multiple, HSPs, as following:

$$S_{cor}^{bit} = \frac{1}{ln(2)} \left(\lambda \sum_{i=1}^{r} S_i^{raw} - ln(kmn) - (r-1)ln(k) + 2ln(g) - ln(r!) \right)$$

where r is the number of HSPs and g is the total gap length between all HSPs. Note that the equation includes the adjustments of Equation 1, thus, S_i^{raw} denotes the initial bit score returned by BLAST for each HSPs.

Protein length correction. (Issue 3) For the analysis in this thesis, another issue arises. The length of the protein sequence limits the maximum of the bit score, two proteins sequences with different lengths have two different maxima. Thus, we calculated the individual maxima for all found proteins and used the percentage of the maxima as the final corrected BLAST bit score.

3 Assessing Adequacy of the Sample Coverage and Annotation in Metagenomics

In this section we consider the problem of inspecting metagenomic sequencing data in terms of covering the original habitat. More specifically, we examine if the sets of features, such as enzymes or species, represent the environmental sample sufficiently.

3.1 Introduction

Reliable classifying and recognizing if a metagenomic sample represents the habitat is of crucial importance for the further analysis of the data set. An incomplete representation either caused by the sequence data or the retrieval of information will limit any additional study. On the other hand, the knowledge that the habitat is well described, by covering the original genomic material multiple times, is beneficial as well. Especially if a metagenomic study is one of the first study focusing on an ecosystem, the number of different species as well as the genomic content is unknown. The design of future studies of the ecosystem can be reasoned upon the outcome of previous sampling, so the sample volume may be adjusted accordingly.

A first approach in assessing the adequacy of the sample coverage can be the analysis of the species diversity. Diversity in ecosystems can be classified into α diversity, β diversity and γ diversity. α diversity is the biodiversity in a certain habitat whereas β diversity measures and compares the diversity between a number of ecosystems. γ diversity described the overall biodiversity over a region, thus usually includes multiple ecosystems [TG12, WGF10].

The method of choice in biodiversity measurement is targeted sequencing of 16S (prokaryotic) or 18S (eukaryotic) rRNA gene. This, for the protein synthesis essential gene, is present in all cells. In extensive database with reference 16S/18S rRNA genes is available with Silva [QPY+13]. A common approach is to estimate the species by clustering rRNA sequences into Operational Taxonomic Units (OTUs) based on sequence similarities [CZC+13]. OTUs can be clustered to different degrees of specificity, corresponding to different levels in phylogeny, such as species, class and genus.

One of many tools to estimate the coverage obtained from sampling based on 16S/18S OTU annotations are rarefaction curves. In a rarefaction curve the number of OTUs is plotted as a function of number of samples, here clones. The curve, which

usually starts with a step increase of additional OTU's, stagnates if no additional OTUs are found with increasing sample size. An illustration of different rarefaction curves is given in 2.

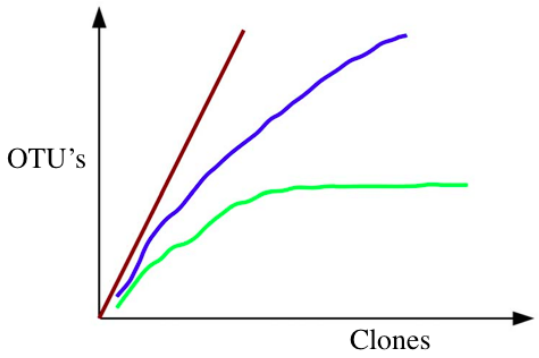


Figure 2: Schematic representation of common rarefaction curves. For the green curve, most or all species are sampled. In the blue curve, the habitat was not exhaustively sampled. Red represents a species rich habitat, only a small fraction was sampled. [WGF10]

Rarefaction curves were for instance used in the analysis of Chrons Disease [PBGLVC$^+$13] and in the analysis of the effect of antibiotics on microbial communities [GHSM12].

There are also several indexes measuring the species diversity, such as Shannon index and Simpson index [PBGLVC$^+$13, AKB09]. Also, newly defined derivations of such indexes arise in the recent years [AKB09]. Those indexes typically are based on count data, therefore the indexes are biased if a the same species is represented by multiple, not exact but very similar copies of 16S/18S rRNA [].

For an analysis of the representation of functional properties similar indexes to the taxonomic representation, like Shannon index, are used [UIL$^+$13].

We will also use several further basic statistical tools in to infer the representation of our habitat. The application fields of these methods are wide spread, thus it is not surprising that they have been used in some way in connection to metagenomics before, for instance in generally designed toolboxes for metagenomic analysis [SH08].

One of these methods are Venn diagrams. John Venn (1834-1923) introduced these as a means of testing the validate of categorical sets. They visualize inclusion and exclusion between sets by a number of intersecting cycles [Bri14].

Another graphical method to summarize qualitative data. A cycle is divided into slices, the size of each representing the proportional number of elements in this

corresponding category [Bri14].

3.2 Methodology

Accuracy and coverage

We estimate the accuracy of the functional and taxonomic annotation and of the coverage of the original genomic context with a modification of the rarefaction curve.

Bootstrapping is a method for testing and estimate the confidence interval for unknown parameters. Bootstrapping randomly subsamples, with replacement, the original data set and calculated the to be estimated value based on the subsamples. The average of all subsamples, as well as the distribution helps estimate the confidence of the initially derived value [Efr03].

We assume a set of metagenomic sequences $S = s_1, ..., s_n$, and sets of features $E = \{e_0, e_1, ..., e_k\}$, where each e is a feature such as an enzymes or a taxon on a taxonomic levels. One features is always reserved to be undefined $e_0 =?$. Each s_i maps to one element in E. We plot the the number of features as a function of the number of the sequences. We generate samples with increment of 2500 sequences. At each of the sampling points, we apply bootstrapping.

Furthermore, the plots will be supported by indicated the maximums $|E|$ of possible γ diversity. Sets of the different metagenomic and genomic samples are taken as estimates γ diversity representations. One line is plotted at the maximum $|E|$ for all sample contigs, another line at the maximum of all reads and a further line at the maximum $|E|$ for all the samples, thus, this maximum is indicating the γ diversity.

Basic comparison of metabolic content

Given feature sets from similar environmental samples we expect considerable intersection and modest relative complements. Logical relationships between small numbers of sets are easily visualized by Venn diagrams. To increase the intuitively, the sets are drawn in proportion to each other.

Quantitative composition of the metagenome sample

So far, we have only consider the features as binary entities without any hierarchical structure or quantity. The relative proportions of OTUs of the data is easy to obtain

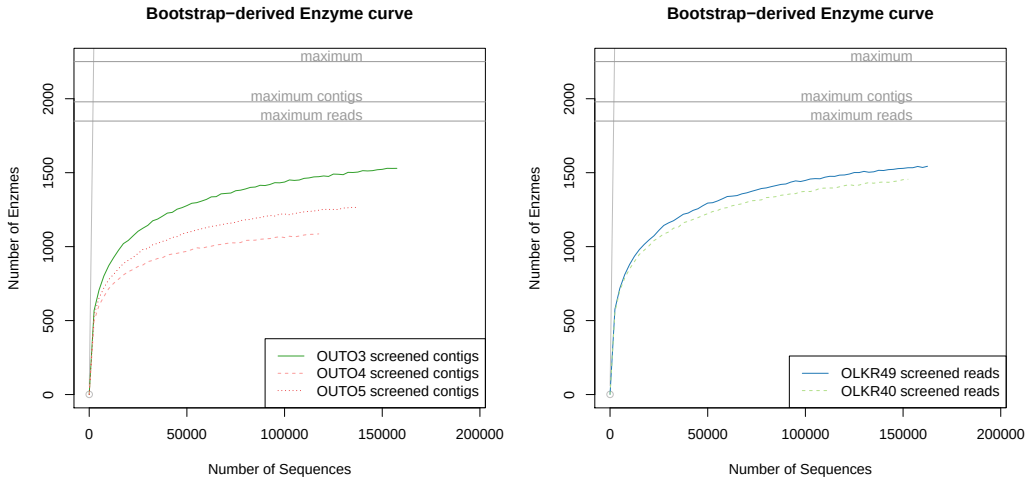
by extracting collections of features instead of sets. In case of the species feature, the hierarchical structure can be derived from taxonomic trees.

In metagenomic samples the species composition is of interest. Generally, it is obtained by a 16S rRNA analysis. 16S rRNA are ideal marker genes to reconstruct phylogenies as they are essential to any organism and consist of highly conserved as well as hypervariable regions.

3.3 Results

We will now apply the previously presented methods on the metagenomic samples and selected reference individual genomes presented in Chapter 2.

First we will consider the accuracy and the coverage of the enzymes which we analyze with the modified rarefaction curves.

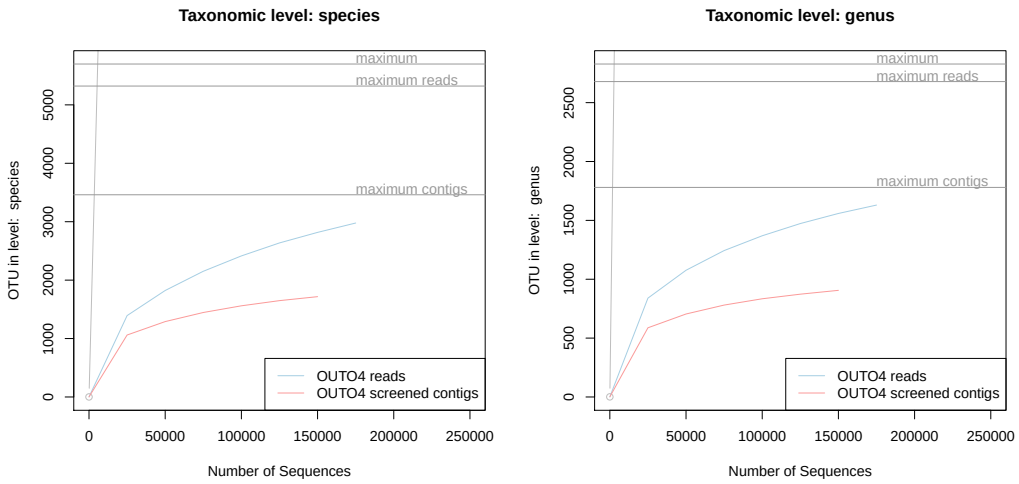


(a) Enzyme of the OUTO samples (b) Enzymes of the OLKR samples

Figure 3: Number of Enzymes in bootstrapped metagenomic samples, shown as a rarefaction curve. Maximum shows the number of enzymes from all data sets, maximum reads, and maximum contigs show the maximum of all data sets, all data sets derived from read data and all data sets derived from available contig data

Figure 3 shows the modified rarefaction curves of the novel OLKR and OUTO samples for the feature sets of Enzymes. We observe that all of the growth of curves stagnates although they do not completely converge. These curves are similar to a curve that represents all of the features as seen in Figure 2 (green curve). The

OLKR samples, which are only available as read sequences, converge slower than the contig samples of OUTO. The overall found enzymes of the OLKR samples and OUTO3 are in a similar range and are higher than the OUTO4 and OUTO5 number of enzymes.



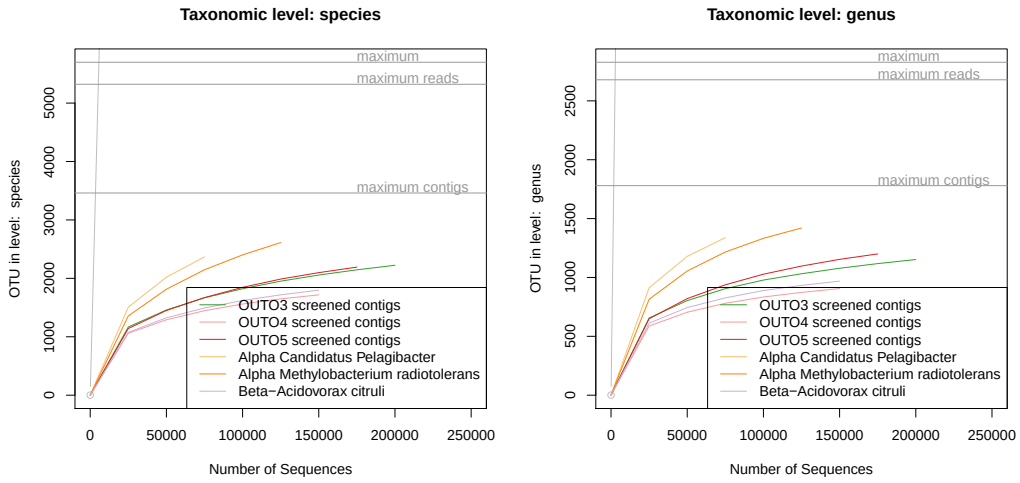
(a) OUTO4, read and contig samples, and selected reference samples for species

(b) OUTO4, read and contig samples, and selected reference samples for genus

Figure 4: Comparison of OUTO 4 read and contig samples, shown with rarefaction curves. Maximum, maximum reads, and maximum contigs show, respectively, the maximum of all data sets, all data sets derived from read data and all data sets derived from available contig data. NOTICE the different scale of (a) and (b)

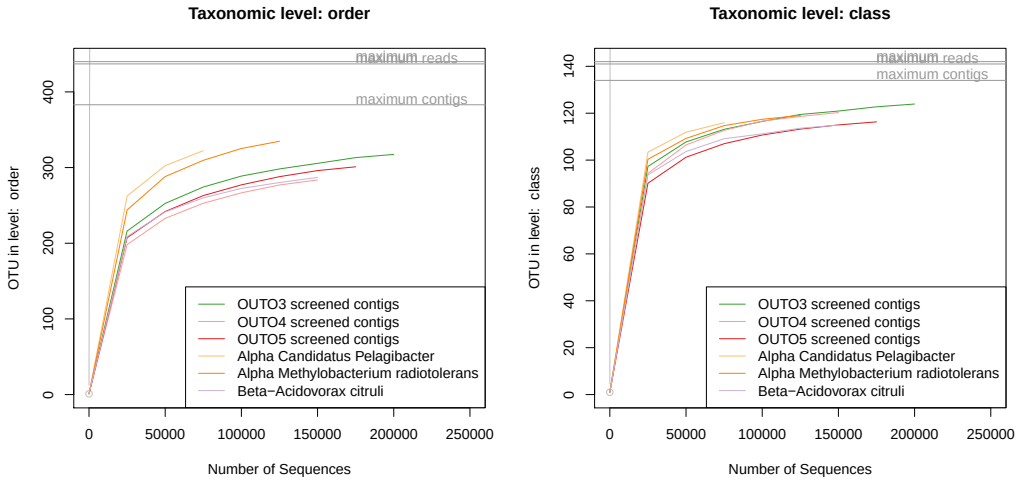
In Figure 4 the rarefaction curve of species and genus of OUTO4 are shown, in each case for the assembled and for the unassembled sequences. Two different taxonomic levels are show consequently the absolute number of OTU's differ. Whereas the contig based curves converge quickly, the read based curves converge slower, and do not reach a plateau. In addition, the overall found OTU's differ greatly between assembled and unassembled data, both on genera and species level of the taxonomic annotation.

Figure 5 shows the rarefaction curve for the sample of OUTO. We shows four different taxonomic levels, species, genus, order and class. The higher the taxonomic level, the more the rarefaction curve converges with increasing sequencing size. This consequently resolves from the decreasing number of different OTU's in each taxonomic level. The class and order level are well represented, whereas the species level is not exhaustively captured by the OUTO samples. The genus level seems to be



(a) Species for OUTO3, 4, 5 and selected reference samples

(b) Genus for OUTO and selected reference samples



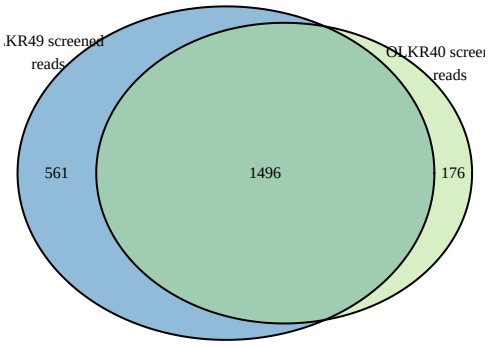
(c) Order for OUTO3,4,5 and selected reference samples

(d) Class for OUTO3,4,5 and selected reference samples

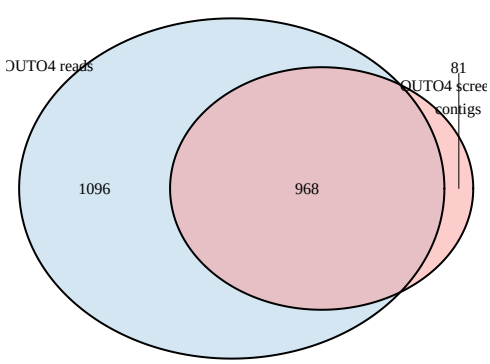
Figure 5: Average number of different taxonomic units in bootstrapped metagenomic samples for OUTO. Maximum, maximum reads, and maximum contigs show, respectively, the maximum of all data sets, all data sets derived from read data and all data sets derived from available contig data

somewhat well represented by the sequence samples. (Notice the difference in scale of the species and genus rarefaction curve.) Overall the different OUTO samples behave similarly.

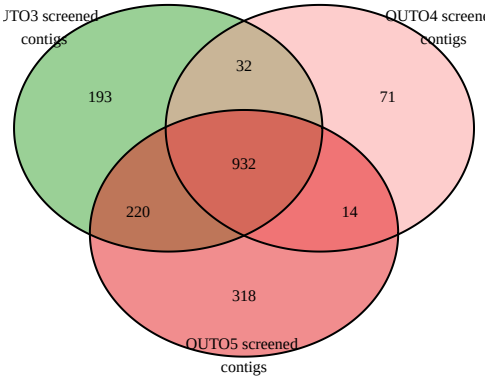
It should be noted that, due to the inclusion of {?} in the feature set, the not mapped sequences map to ? and thus shift the values of the y-axis by 1.



(a) OLKR49 and OLKR40



(b) OUTO4 reads and contigs

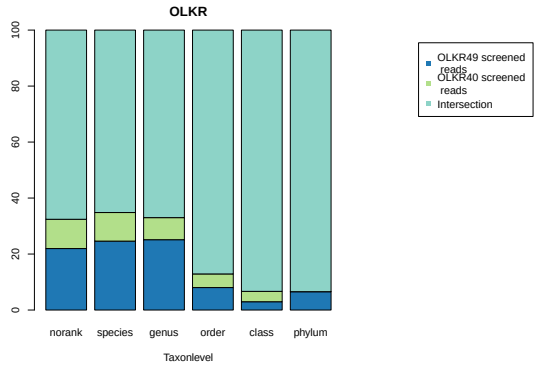


(c) OUTO3,4,5

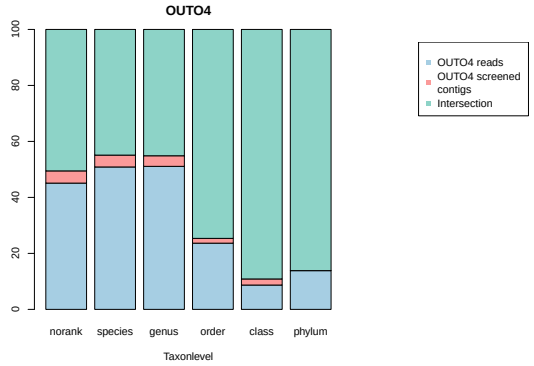
Figure 6: Number of genus and their overlap of different sample combinations.

In Figure 6 Venn diagrams for the genus level are shown. The Venn diagram in 6a shows the difference between the the two samples of OLKR, OLKR49 and OLKR40. Although both samples contain mutually exclusive genera, the majority of the genera is the same. The Venn diagram in 6c shows the differences between the three OUTO samples. Each sample has its own exclusive set of genera, OUTO5 has the largest number of unique genera. In addition, the sample OUTO5 contains many genera also found in OUTO3 but not in OUTO4. The differences in annotated genera number between OUTO 4 assembled and unassembled become obvious in Venn diagram 6b

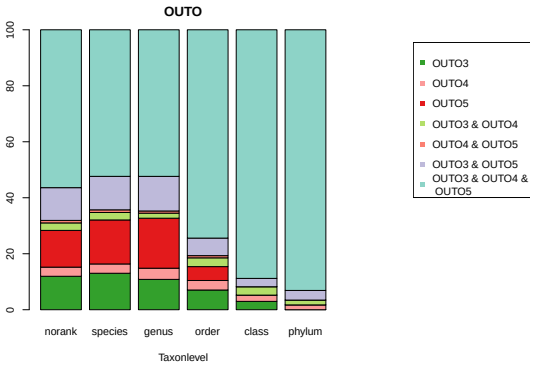
In Figure 7 summaries of the different possible Venn diagrams for the three setups for the taxonomic annotation. In absolute numbers, the higher the taxonomic level the bigger the intersection between the different samples, and the smaller the number of unique OTUs in a sample. The summary shows the individual contributions of the samples as percentages, thus, generating a normalization for every taxonomic level. In Figure 7a the two OLKR samples are examined. The OLKR49 sample contains more unique OTU's than OLKR40 on all levels. On the phyla level OLKR40 can be



(a) OLKR49 and OLKR40



(b) OUTO4 reads and contigs



(c) OUTO3,4,5

Figure 7: Summary of logical relationships represented by Venn diagrams.

seen as merely a subset of OLKR49. In Figure 7b the assembled and unassembled OUTO4 samples are shown. As already seen in previous investigations, the read data continuously annotates a greater amount of OTU's. On the lower taxonomic levels, norank[1], species and genus the read sample consists of twice as much OTU's than the contig sample. In Figure 7c the three OUTO samples are examined. A

[1]This level the same as a strain, a subtype of microorganisms, a genetic variant.

clear core of OTUs for all three samples on all taxonomic levels is observed. OUTO5 and OUTO3 have a bigger set of unique OTUs than OUTO4.

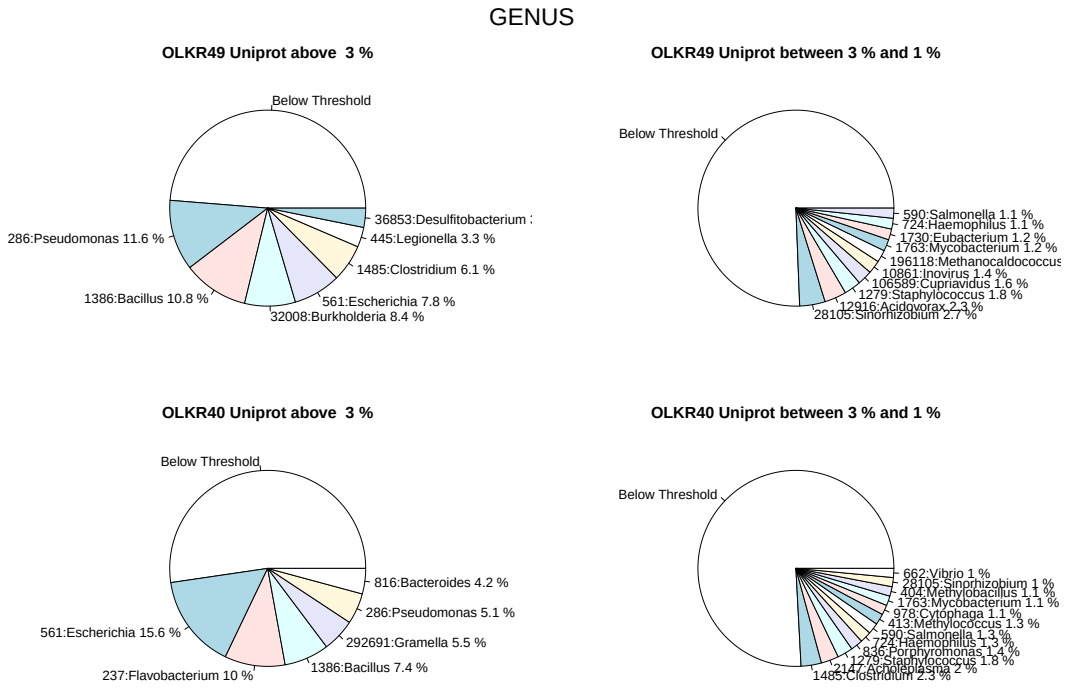


Figure 8: Genus composition for the OLKR40 and OLKR49 samples

Figure 8 shows the OLKR40 and OLKR49 composition of the genus taxon. Major genera, bacilli, escherichia and so on, are found in both samples. However, genera only contributing to a smaller extend to the community are usually not found in both samples. The percentage of the contribution of each genus vary greatly between the closely related habitats.

3.4 Discussion and Conclusion

Advantages of Assembled Samples We were able to observe multiple times the effects the assembly of the data had on the outcome of the results. Rarefaction curves for read data reach higher maxima and converge slower. Venn diagrams support the observations that the read data annotates to a greater number of OTU's or enzymes. I can be assumed that, because the read data is less specific in their annotation than the contigs, the read data annotates randomly to closely related

features, either OTU or Enzymes. Thus, the assembly of the data sets is important. Also, it does infer for the OLKR samples a cautious further analysis.

Enzyme Representation. The enzymes are well represented by our sequencing set as shown by the rarefaction curves in Figure 3. The maxima for a sample of found enzymes as well as the speed of convergence depends on the possibility to assembly the data, thus, explaining the slight difference between the OLKR and OUTO samples. Although the overall found enzymes in OUTO3 is similar to the number of enzymes found in OLKR it is reasonable to assume that OUTO3 is richer in enzymes, because the functional annotation was based on the more specific contig data sets.

Taxonomic Representation Based on the rarefaction curves and the Venn diagrams, genus is a somewhat reasonable represented taxonomic level in all the samples. Closer inspection of the genus level in the OLKR samples reveals the contributing genera and the proportion. Domain knowledge lets us assume that the proportions might not be correct, however, generally the found genera are reasonable. A analysis of actual 16S/18S rRNA would greatly improve this aspect of the studied metagenome. It could also serve as a reference for correctness of the proportional composition.

4 Pattern Detection with Principal Component Analysis

In this chapter we consider the problem of mining for characteristics differences between the metabolic content of metagenomic samples by means of principal component analysis (PCA) [Pea01].

4.1 Introduction

PCA is a technique of multivariant statistics, a set of processes in which multiple statistical variables are analyzed. The goal of these statistics is the detection of dependency structures between variables. Multivariant statistics are divided into inductive statistics, in which the data is fit to a predefined structure, and into exploitative statistics, in which the structure is attempted to be extracted from the data. In inductive statisics, we assume to already know the underlying structure, but a structure may be incorrect and therefore lead to overfitting. Explorative statics on the other hand is trying the more difficult task to determine unknown structures.

Both staticial fields, inductive and explorative, can be subdivided into typical representants and main methods. Inductive statiscs is mostly represented by regression analysis, which in turn can be subdivided into 3 main methods: Lasso, elastic net and ridge regression [WPT+11]. Explorative statistics has three main approaches. A first approach is the reduction of many variables to latent constructs, such as done with PCA, correspondence analysis and factor analysis. Other approach in explorative statistics include the clustering to groups of observations and multidimensional scaling.

For the scope of this thesis, we will in the following reduce the wide field of multivariant analysis and focus here on PCA used in metagenomics.

PCA is a technique with multiple application fields; (1) dimensional reduction, (2) lossy data compression, (3) feature extraction, and (4) data visualization [Jol86]. There exist two definitions of PCA, both resulting in the same algorithm. Already in 1901 Pearson [Pea01] defined PCA as a linear projection that minimizes an average projection cost. The average projections cost is the mean squared distance between the data points and their projections. In 1933, Hotelling defined the same idea as the orthogonal projection of the data onto a lower dimensional linear space, while ensuring that the variance of the projected data is maximized. We thus have

to formulations, a minimum-error formulation by Pearson and a maximum variance formulation by Hotelling [Jol86]. PCA has been a successful tool for two main reasons. Firstly, the principal components sequentially hold the maximum variability. Therefore, the first principal component holds the most information about the variance in the data, and only the n-first principal components that satisfies a user-defined variance requirement have to be considered. Also, the principal components are uncorrelated and can be interpreted individually [ZHT04].

Furthermore, there are several variances of the classic PCA as mentioned above. For instance, it is possible to express PCA as a probabilistic latent variable model which is dissolved to the maximum likelihood solution. Some advantages of this definition are, that it can deal with missing data and that it can be trained with the Expectation-Maximization-Algorithms (EM-Algorithms).

A further variance from the classical PCA is Sparse PCA. In comparison to the maximum variance formulations by Hotelling, spare PCA only hold up to explain most of the variance of the given data. An approach by Zou and Hastie [ZHT04] uses lasso or elastic net constraint to produce modified principal components with sparse loading. This advantages in so far, that the principal component are no longer forced to be linear combination of all the original variables, also, the loading of each principal component is less likely to be non zero, thus allowing for an easier interpretation of the results.

PCA as a technique is used by a number of metagenomic analysis software. For example, PCA is implemented in the web server for comparative metagenomics METAGENassist [AXL+12]. The samples are displayed on a two-dimensional cluster/scatter plot by selecting two of the principal components. PhyloSift [DJL+14] offers a software pipeline for the phylogenetic analysis of genomes and metagenomes. Here, the comparison between sample phylogenies is done with edge PCA [ME13], which depends on a reference phylogenetic tree, called phylogenetic placements. Edge PCA analyzes a matrix where rows list each sample, columns correspond to edges in the reference phylogeny and the individual matrix cells hold the difference in placed sequence probability masses on either side of the edge. The standard dimensionality-reduction of PCA is applied to find the contribution the corresponding reference phylogeny holds for the variation among the samples in that dimension.

4.2 Methodology

In this chapter we compare the metabolic, not the taxonomic, content of the different samples. In order to do so, we use the publicly available metagenomic and genomic samples described in Chapter 2 as reference and apply standard PCA as well as modification, Sparse PCA. Previous analysis by biologists hinted at similarities in the metabolic content to fungi and/or archaea. Thus, they were included in the genomic reference set. We base our analysis on the Sample $\times$ Enzyme matrix containing the corrected BLAST bit score and to a binarization of the same matrix.

PCA Principal Component Analysis (PCA) was first described by K.Peason in 1901 [Pea01] as a method in physics, statistics and biology to represent points in the plane or other, lower dimensional spaces. PCA transforms the variables into linear combinations, principal components, that describe the maximum variance in the data.

Two approaches of transformation are commonly used. Either a singular value decomposition (SVD) of the original data matrix or an eigenvalue decomposition of the covariance matrix of the original values is performed [ZHT04, dEJL07]. Here, we used the latter approach. Based on the eigenvalue for each eigenvector we can sort the PCs and calculate the variance of the original data they explain.

Sparse PCA (SPCA). In standard PCA, the PCs are linear combinations of all variables. Consequently all weights, also called loadings, of the PCs are usually non-zero. Therefore, in our application, all enzymes are assigned a loading for each PC, and again all enzymes are used to describe each PC. This is a disadvantage because each PC has an underlying biological interpretation [dEJL07]. So the enzymes which contribute to the first PC are associated with the most variance in between our samples. An approach which circumvents this disadvantage is Sparse PCA [ZHT04], which interprets PCA as a regression optimization problem. Regression analysis tries to describe a dependent variable with multiple independent variables. The lasso can be integrated into the regression optimization. Lasso is a variable selection technique which simultaneously produces accurate and sparse models by penalizing based on the difference between the actual values and the values defined by the regression. Typically, the least square penalty term is used.

The here presented sparse PCA was calculated with the 'elasticnet' R package [ZHT04]. The non-zero loadings of each PC were limited to 40.

4.3 Results

In this section, we show a selection of PCA results which focuses on the following aspects:

1. Comparison of PCA and SPCA.
2. Comparison of corrected BLAST bit score and binary data.
3. First four PCs for OLKR.
4. First four PCs for OUTO.

Table 6 list the corresponding samples for all following PCA plots which contain number annotation.

Number	Sample
1	OLKR40
2	OLKR49
3	OUTO3
4	OUTO4
5	OUTO5
6	OUTO6
7	GW
8	MAAOC
18	Alpha Candidatus Pelagibacter ubique HTCC1062
19	Alpha Methylobacterium radiotolerans JCM 2831
20	Beta-Acidovorax citruli
21	Beta-Acidovorax ebreus TPSY
22	Beta-Methylotenera versatilis 301
23	Beta-Methylotenera mobilis JLW8
24	Gamma Pseudomonas fluorescens F113
25	Gamma Pseudomonas putida F1
26	Gamma Methylococcus capsulatus str. Bath
27	Gamma Stenotrophomonas maltophilia K279a
28	Delta Desulfobulbus propionicus DSM 2032
29	Delta Desulfovibrio aespoeensis Aspo-2
30	Delta Geobacter sulfurreducens KN400
31	Epsil Sulfurimonas denitrificans DSM 1251
32	Epsil Sulfuricurvum kujiense DSM 16994
33	Epsil Syntrophobacter fumaroxidans MPOB
34	Flavo Flavobacterium johnsoniae UW101
35	Flavo Flavobacterium psychrophilum JIP02/86
36	Chlor Anaerolinea thermophila UNI-1
37	Chlor Chloroflexus aurantiacus J-10-fl
38	Firmi Syntrophomonas wolfei
39	Firmi Veillonella parvula DSM 2008
40	Firmi Desulfotomaculum acetoxidans DSM 771
41	Firmi Desulfitobacterium dehalogenans ATCC 51507
42	Firmi Desulfotomaculum reducens
43	Firmi Desulfosporosinus acidiphilus SJ4
44	Archaea Methanothermobacter thermautotrophicus str. Delta H chromosome
45	Archaea Methanosarcina acetivorans C2A chromosome
46	Archaea Methanosaeta concilii GP6 chromosome
47	Archaea Thermoplasma volcanium GSS1 chromosome
48	Fungi cryptococcus neoformans grubii h99 2 contigs
49	Fungi neurospora crassa or74a finished 10 contigs
50	Fungi ustilago maydis 1 contigs

Table 6: Universal number of individual samples in all PCA plots.

Comparison of PCA and SPCA. We will here now examine the different results for PCA and SPCA, applied to the same data sets.

First, Table 7 shows the average running time for PCA and SPCA, applied repeatedly to our data sets. The standard PCA shows a significantly better time performance in practice.

In Figure 9 we compare the outcome for SPCA and PCA of OUTO together with two metagenomes as well as publicly available genomes. Both plots show the first against the second PC. Overall, the positioning of the samples on the plots are very similar. Slight changes of positioning are observed for 11 and 12, Pseudomonas. The axis ranges differ, due to the difference in transformation. The important directions are maintained.

Comparison of corrected BLAST bit score and binary data. In general, our lower quality metagenomic samples obtain lower corrected BLAST bit scores. Here, we compare the corrected BLAST bit score PCA results, based on binary data just indicating whether or not we detected an enzyme.

In Figure 10 the PCA plots for PC1 against PC4 are given based on corrected BLAST bit score and on the binary data. The plots differ, the richer metagenomic OUTO samples are put to an extreme. The PC4 only accounts for around 56% of the variance, however, this was the case for most of the plots (`http://www.cs.helsinki.fi/group/urenzyme/deepfun/OLKI/SPCA_Bacteria_Binary/`). It appears, that the metagenomic samples, which in general hit more, diverse enzymes, are now overly emphazised.

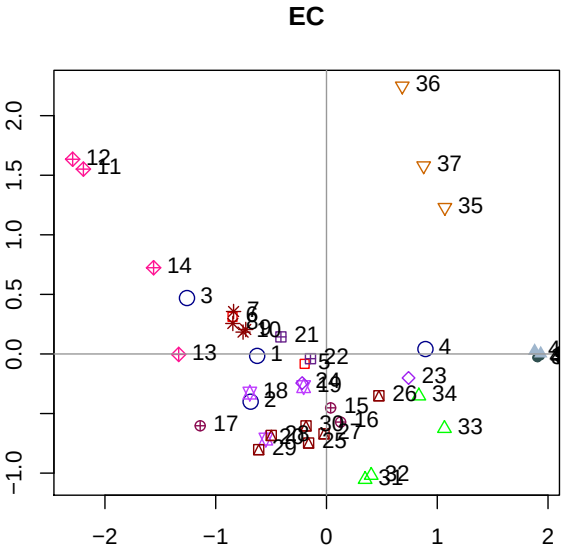
We now examine both, the OUTO and the OLKR samples further by exploring the first 4 PCs of both SPCA analyses. Figure 11 shows several PCA plots for the OLKR analysis together with the reference metagenomes and genomes. The first four PC together make up around 80 % of the variance. The fungi and archaea are usually situated further away from the bacteria. The OLKR samples lay within the bacteria.

	User	System	Elapsed
Sparse PCA	1144.9723	0.1542	1145.9709
PCA	17.2519	0.0586	17.3712

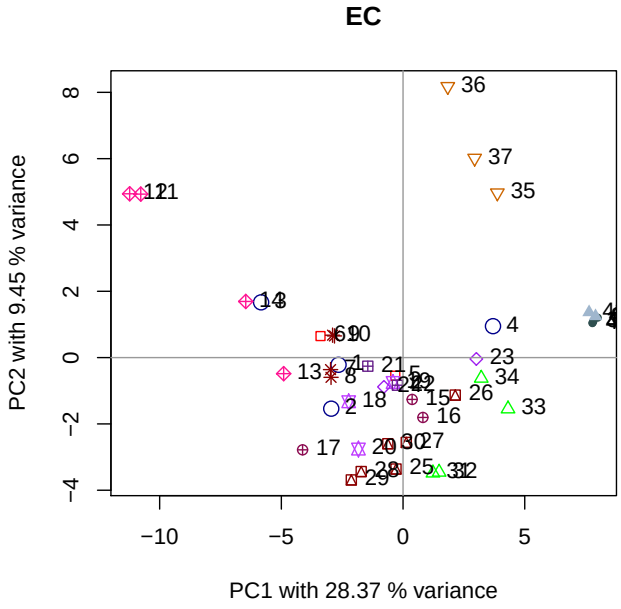
Table 7: Average running time (seconds) for Sparse PCA and PCA. Each method was applied 10 times to 5 different sized matrices.

The two Gamma Pseudomonas bacteria 11 and 12 are usually somewhat separated from the other samples. Pseudomonas is an organism appearing in various habitats. A further investigation of the enzymes we detected for these bacteria revealed, that they seem to hold enzymes none of the other bacteria hold. Therefore, the unique placement of these is reasonable.

In Figure 12 the first four PC of the OUTO analysis are shown. Together they around 56% of the variance in the sample. The OUTO samples are richer in their metabolic content and they have more enzyme annotations. Therefore, the generated Sample $\times$ Enzyme matrix is less sparse. Nevertheless, similar to the OLKR samples, the fungi and archaea are separated from the bacteria and the OUTO samples. Pseudomonas is again separated from the rest of the bacteria.



(a) Sparse PCA. PC 1 against PC2, for



(b) PCA. PC 1 against PC2

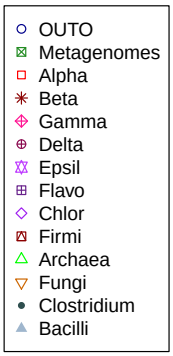
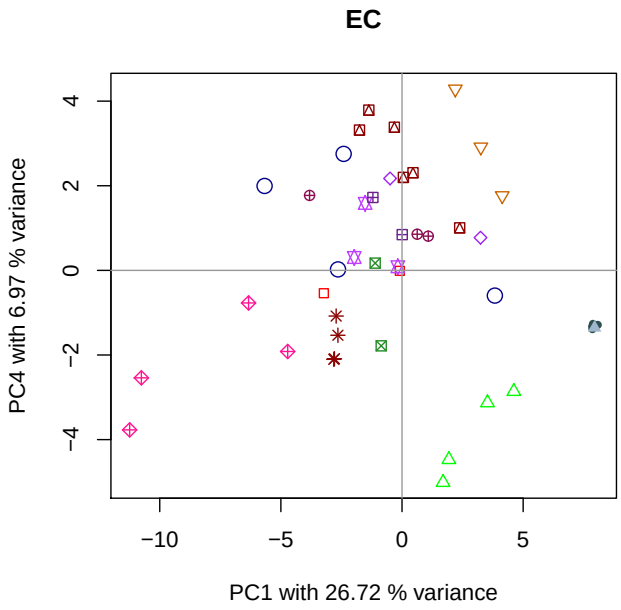
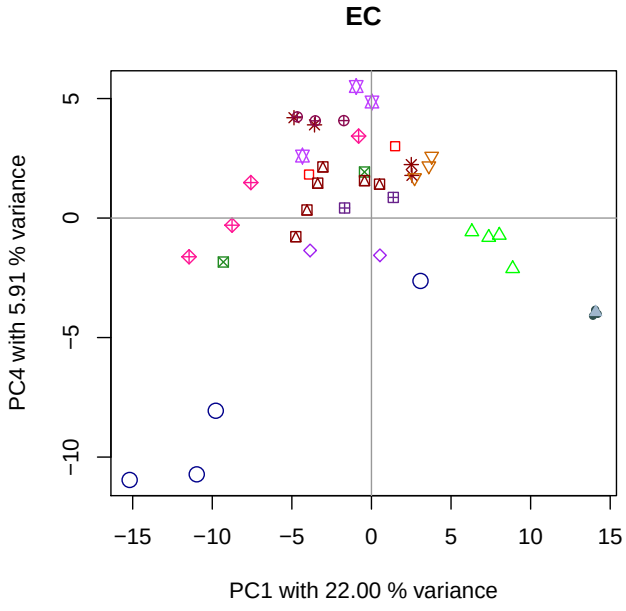


Figure 9: Comparison of PCA and Sparse PCA results. Displayed are the OUTO samples, public metagenome samples and the public genomes.



(a) Corrected BLAST bit score results, PC 1 against PC2



(b) Binary results, PC 1 against PC2

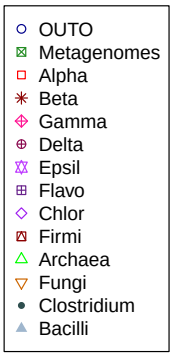
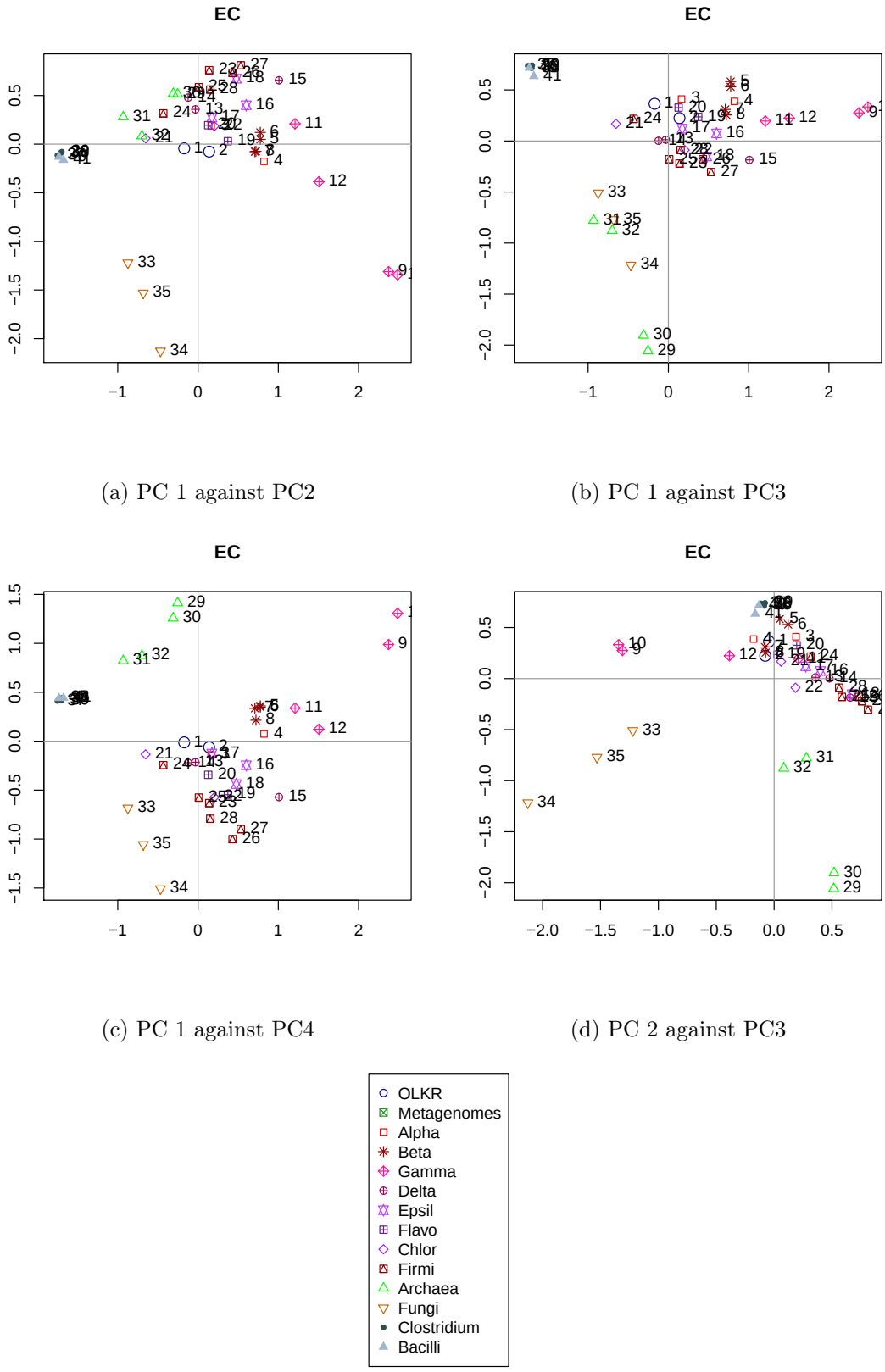


Figure 10: PCA of binary and corrected Blast score results for OUTO.



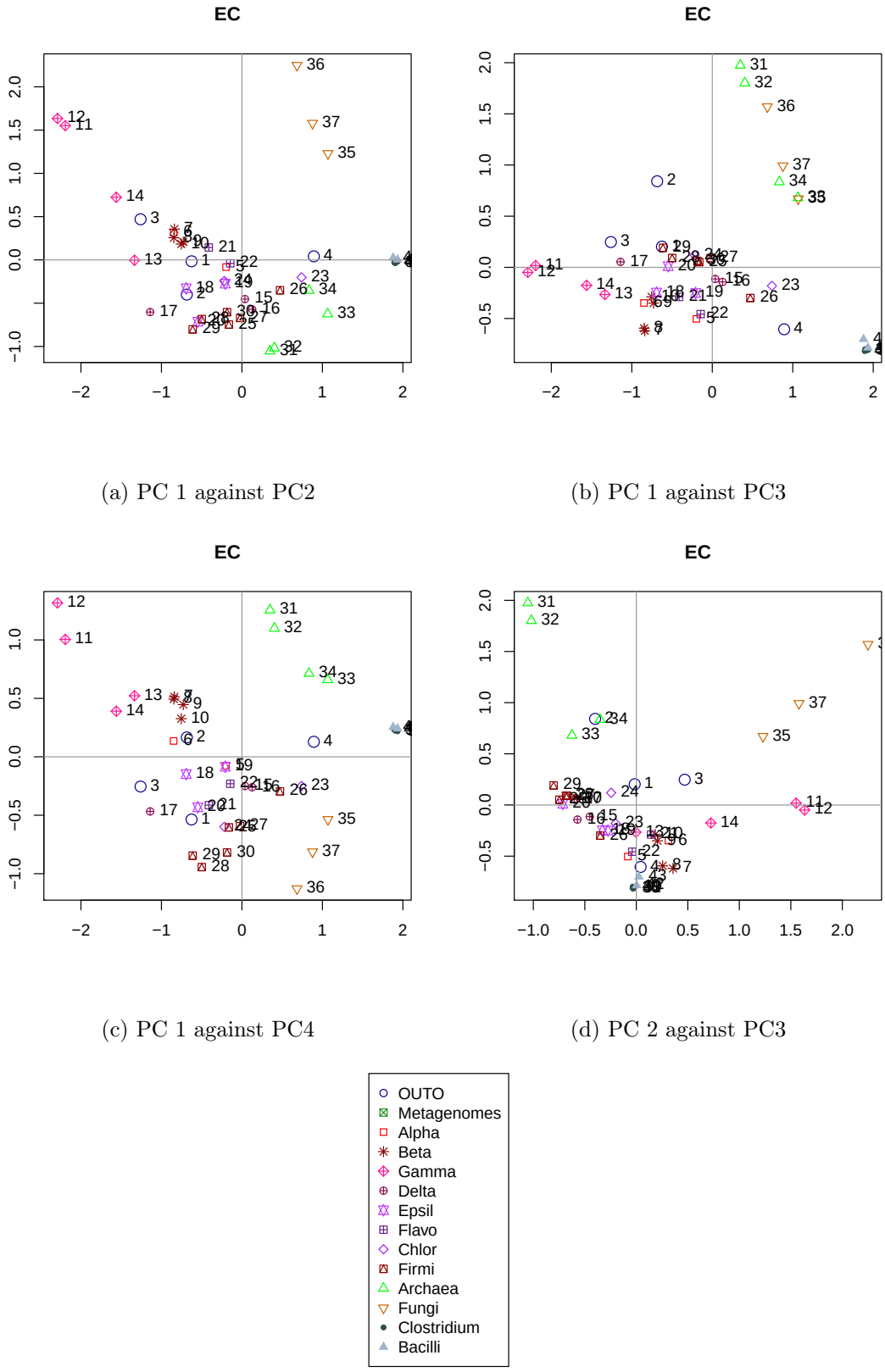
(a) PC 1 against PC2

(b) PC 1 against PC3

(c) PC 1 against PC4

(d) PC 2 against PC3

Figure 11: Sparse Principal Component Analysis for OLKR.



(a) PC 1 against PC2

(b) PC 1 against PC3

(c) PC 1 against PC4

(d) PC 2 against PC3

Figure 12: Sparse Principal Component Analysis for OUTO.

4.4 Discussion

The differences in SPCA and PCA results are small. The minor constraint relaxations of Sparse PCA affect our results to a minor extent. The gain in intuition of interpretation is great. Therefore the more time-consuming SPCA is worthwhile for this kind of analysis.

Using binary data matrices has not proven to generate fruitful results. The differences between metagenomic and genomic samples were highlighted, instead of highlighting differences between their metabolic content. In general PCA/SPCA was sensitive to differences in the sample preparation. Therefore, analysis of similarly generated samples should be favored.

Neither the OLKR nor the OUTO samples have shown great similarities in terms of the metabolic content to fungi or archaea. Thus, the initial assumption of fungi or archaea could not be independently confirmed. More specific analysis, maybe with a higher quality data sets, might still hold up to the initial assumption; however, this is not expected.

5 Metabolic Pathway Visualization

In this section, a visualization methodology of the functional annotation, the enzymes, is proposed.

5.1 Introduction

The Kyoto Encyclopedia of Genes and Genomes (KEGG) is a reference knowledge database linking genomes to biological systems and environments. Since it was first introduced in 1995 by the Kanehisa Laboratories, 16 different databases have been established. The databases can be classified into four major group:

1. systems,
2. genomic,
3. chemical,
4. and health.

The reliability of the information given is held high by basing the presented information on published and verified experimental results. One of the project's first goals was the KEGG pathway database which holds manually drawn pathway maps visualizing for instance metabolic pathways [Kan96].

As recently stated in "Data, information, knowledge and principle: back to metabolism in KEGG" (2014) [KGS+14] the future outlook for the KEGG pathway maps is a switch from Enzyme Commission number (EC) annotations to KEGG's own classification system, KEGG orthologs (KO). KEGG orthologs are manually defined and context-dependent groups of orthologs genes. Ortholog genes are genes in different species which originate from the same gene in a common ancestor. Increasing discrepancy between EC number classification and pathways has made this switch of annotation necessary. Not all pathways have an associated EC number and a lack of sequence information of many enzymes has lead to only half of the pathways of the KEGG pathways maps to be annotated with EC numbers [KGS+14]. A web-based server, KAAS [MIO+07], has been developed to support future genome and gene mapping to the K number of KO.

KEGG pathway maps are widely used in metabolic capacity visualization. Hence they have been integrated into metagenomic analysis pipelines such as MG-Rast

[GWW+10] and IMG-M [MCC+12], an extension of IMG [MCP+12], to accommodate metagenomic data.

Vanted [RJH+12, JKS06] is a general support software for multi-omics data sets. The software supports dynamic networks and allows for automatic integration of information from various databases. Initially, it largely depended on the KEGG database, integrating the metabolic information via the KGLM file format [KAG+08], but it now incorporates various databases and their formats. The visualizations in the Vanted software are dynamic, and the user can manually edit the network.

A competing metabolic pathway database to KEGG is Metacyc [MCP+12]. Metacyc pathways are, similar to KEGG, experimentally derived and are widely used as a reference. Metacyc visualization is supported by the 'Pathway Tool' of the Biocyc database. A recent comparision by Altman *et* al. [ATK+13] between KEGG and Metacyc showed, that the KEGG database contains significantly more metabolites, also called compounds, but the Metacyc database contained significantly more reactions, thus pathways.

In the course of this thesis we show examples of KEGG pathway maps uses for the metagenomic samples presented in Chapter 2. The full selection of KEGG pathway maps is available online at `http://www.cs.helsinki.fi/group/urenzyme/deepfun/colorpathway_screened/`. It is also possible to see the colored versions on the maps on the original KEGG webpage, allowing the user to interactively retrieve more detailed information about the elements on the maps, such as the single pathways and the compounds.

5.2 Methodology

We superimpose the corrected BLAST bit score and the count data separately on 382 KEGG metabolism pathway maps. Primarily, EC number annotations were used. When no EC number was available, the KO number was taken into account. Annotation was based on alignment of the sequence reads or contigs to Uniprot entries, as described in in chapter 2.

A coloring scheme from blue to red, ranging from 0 to the highest corrected BLAST bit score or count of the sample was calculated. The color coding of all maps of one sample is universal; however, it is not comparable between samples. If no Blast bit score or count is assigned to a reactions on the map, it is displayed in grey. Reactions in white indicate that it was not possible to color these due to the

ongoing annotation migration process of KEGG. Yellow reactions are not catalyzed by enzymes and occur spontaneously.

For each map either the corrected Blast bit values or the counts of one sample were displayed. In addition to a community map showing the whole samples annotation, the three species with the highest overall reaction count per sample were selected and maps just displaying the corrected Blast bit score or count for these species were generated.

The KEGG REST API was used to automatically retrieve the maps from the online web interface.

5.3 Results

For the sample OUTO3 the six species with the most EC and KO annotations were: Bacillus cerus (Taxon id: 1396), Streptococcus pyogenes (Taxon id: 1314), Streptococcus pneumonia (Taxon id: 1313), Clostridium botulium (Taxon id: 1491), Bacillus subtilis (Taxon id: 1423), and Lysterina monocytogenes (Taxon id: 1639). The overall highest corrected BLAST score for one EC or KO in the complete sample was 0.857 and the overall highest number of hits to one EC or KO was 184.

As an example we will inspect the basic metabolism pathway of the citrate cycle for sample OUTO3 of the whole community as well as three species specific maps. The aerobic citrate cycle poses as a possible energy supply source for bacteria [Lev11]. In Figure 13 the topmost map shows that most of the pathways of the citrate cycle are present in our overall OUTO3 community. The enzymes 4.1.3.6, 1.1.1.41, 1.2.7.3 are not present in our community, nevertheless the compounds of these pathways are connected indirectly. However, we have no evidence of the OUTO3 community to transfer the compound *Fumarate* to the compound *Oxaloactate*, or the other way around.

Moving on to the species specific map of Bacillus Cereus we notice that only three pathways are covered. As described in the Biomodels database [EMB14], the genome sequence of Bacillus cereus seems does lack genes for the full cycle, but contain genes for specific segments of the cycle. However, the samples do not cover all of the pathways associated with Bacillus cereus. The Streptococcus pyogenes pathway map shows no colored pathways. A doctoral thesis [Lev11] focusing on this species has shown that the citrate cycle is mostly missing in Streptococcus pyogenes. Although it is capable of citrate uptake, energy is produced thought glycolysis and pyruvate metabolism in Streptococcus.

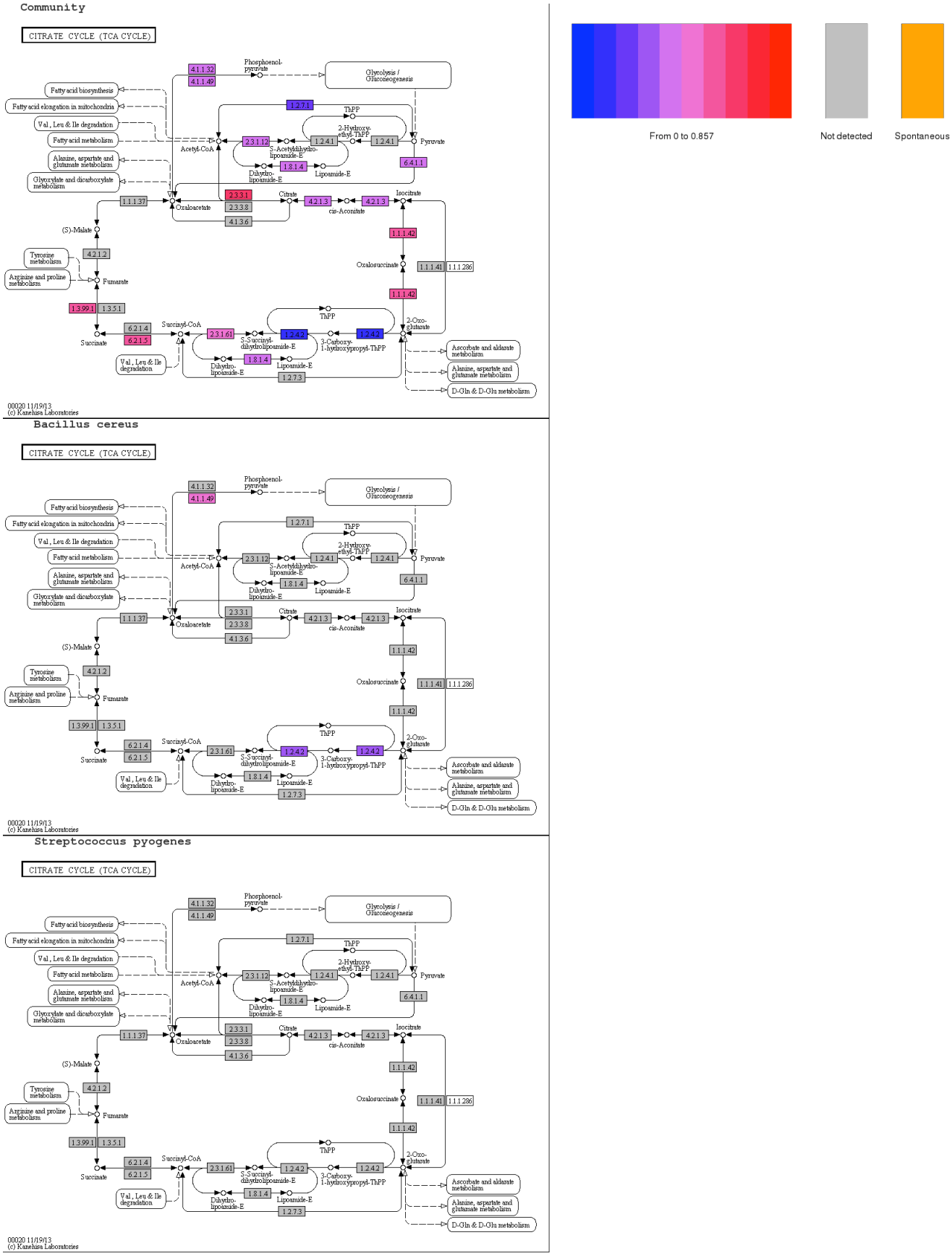


Figure 13: Citrate cycle, OUTO3 sample

5.4 Discussion

The presented KEGG maps enabled us to visualize metabolic pathway findings of sequencing data in a way that assists biological interpretation. The maps show models that are highly reliable because they are based upon published, peer-reviewed results.

A drawback is that the tedious process of drawing the maps causes are slightly outdated and any partly superficial pathway maps. Any pathway defined is just an abstraction and will be superficial. Furthermore, the layout and arrangement of the maps leaves little flexibility in the selection and arrangement of metabolites and pathways to the user. At the same time, the strict, unalterable maps are easily recognized and make for a quick orientation.

Due to the currently on-going annotation transition of the KEGG maps, it can not be ensured that all annotations of the here presented maps are colored. Thus white, uncolored pathways are present in the maps. After the completion of the migration process by KEGG the KO annotation the here presented method can easily be adapted to the change. Further improvements of the KO annotations, using such tools as KAAS, could be considered.

In the shown citrate cycle map the full citrate cycle was not recovered based upon our OUTO3 sample. One possible explanation is that the citrate cycle is actually not present in the community. Another possibility is that the recovery of the data was incomplete either at the sampling, sequencing or functional annotation step. However, it become obvious that individual organisms are only able to execute part of the cycle and that only the full community is capable of full citrate cycle.

It is unlikely that Streptococci pyogenes is in any of our data samples as it usually colonizes skin and tonsils of humans, causing different infectious diseases. Assuming no contamination during the sampling and sequencing procedure, we could conclude that OUTO3 holds a species, or species group, which is most similar to Streptococci pyogenes and either is a novel or a not in uniprot presented species. However, another conclusion could be that we simply observed many random similarities to the well explored S.pyogenes, that would suggest that we need further techniques to avoid spurious alignments.

6 Taxonomic Distribution of Pathways

In this Chapter we want to analyze how the pathways are distributed between the different taxons in our samples. We apply a novel strategy of kernel methods and k-mediods clustering to achieve this goal.

6.1 Introduction

So far, we have characterized the environmental habitat of deep biosphere bedrock by the taxonomic distribution and the metabolic content. Advancing on this now, we combine these two fields and determine which taxon are involved in which metabolic processes.

At present, such analysis focus on specific, essential pathway components. So is for instance an extensive review available concentrating on the distribution of the six different CO2 fixation pathways in autotrophs of extreme biotopes [MCRFAS12]. The presented findings are mostly based on experimental results. A more computational approach, this time focusing on the nitrogen fixation, predicts possible capabilities of this biochemical pathway in species not yet known to hold nitrogen fixation means [DSFM$^+$12]. Nitrogen fixation is difficult to experimentally determine, but in the paper a novel set of genes is proposed as minimum criterion and extensive searches for these in metabolic genomes were constructed.

Another study focuses on the geographic distribution sulfur oxidation and OTU's between terrestrial sulfidic spring [HE13]. This study, based on statistical analysis, such as canonical correspondence analysis, was also able to give evidence of niche space for bacteria oxidizing reduced sulfur compounds.

Here, we present a novel computational strategy for the taxonomic distribution of pathways based on kernel methods as well as clustering approaches. Kernelized clustering approaches have been used before in metagenomic analysis. TACOA [DKG$^+$09] is an approach for a taxonomic classification based on kernelized k-nearest neighbor clustering. K-nearest neighbor (k-NN) is a comonly used classifying approach, which assigns objects based on a majoritiy vote of the k nearest neighbours. Kernels are methods that transfer data into a different, higher dimensional space in which a previously impossible calculation or clustering is possible. A well know representative of these method is the support vector machine, SVM. TACOA now circumvents a major drawback of kNN, the 'curse of dimenionaltiy' in high di-

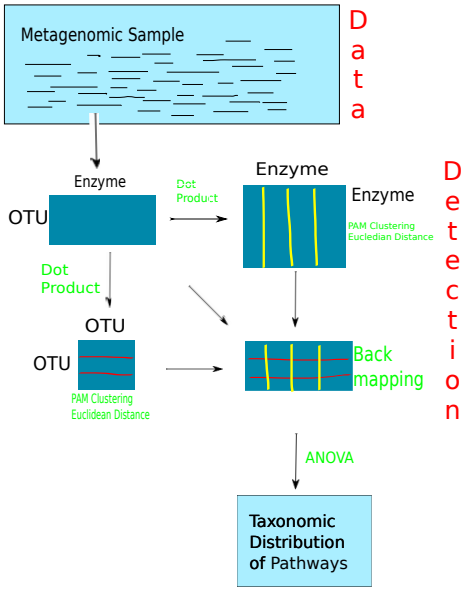


Figure 14: Outline of the novel strategy for taxonomic distribution of pathways mining. An OTU × Enzymes matrix is generated based on the metagenomic samples. The identification is performed in two steps. A kernel method, here the dot product, transforms the original data into OTU × OTU and Enzyme × Enzymes matrices. Partitioning around mediods, PAM, clustering is performed on these, and a back mapping, using the original data, creates a two cluster view of the matrix.

mensional data by applying a Gaussian kernel. By smoothing the data with the kernel, the complete reference database is considered for the clustering instead of strict neighborhoods. The general problem TACOA tackles, supervised phylogenetic clustering, is a sophisticated solution to the phylogenetic annotation performed in Chapter 2, based on homology identified with BLAST. However, the analysis of TACOA is limited to on sequencing methods generating longer sequencing reads.

Our novel stategy for taxonomic distribution of pathways determination is a kernelized k-medoids clustering approach.

6.2 Methodology

Here we now outline a novel strategy for the the detection of taxonomic distributions of pathways. An overview of the outline is shown in Figure 14. Based on the metagenomic sample, an OTU × Enzyme matrix is generated. The approach used

is described in Chapter 2, but other annotation approaches can be used, as well. An inner product kernel is used to transfer this matrix into two different spaces, one pairwise comparing the enzyme profiles of the OTU's (OTU $\times$ OTU), and one comparing the OTU profiles of the enzymes (Enzymes $\times$ Enzymes).

Given the objects $p = (p_1, p_2, ..., p_n)$ and $q = (q_1, q_2, ..., q_n)$ the dot product $k(p, q)$ is defined as

$$k(p, q) = \langle p, q \rangle = \sum_{i=1}^{n} p_i q_i$$

After this, a partitioning around mediods, PAM, clustering is applied. The PAM clustering, also k-median, clusters the set of objects into k classes based on minimizing distances between the objects in a class. Each class is represented by the object with the overall smallest distances, the mediod. The mediod represents supports an intuitive interpretation of the clusters. We used the PAM clustering [KR90] implemented in the R package 'cluster'. The clustering method used the Euclidean distance, which is defined as:

$$d(p, q) = \sqrt{\sum_{i=1}^{n} (p_i - q_i)^2}$$

In order to verify the goodness of the clustering, silhouette [Rou87] was used.

The Silhouette coefficient is a measument for cluster cohesion and separation first introduced by P.J. Rousseeuw [Rou87]. accurary of individual object classification, the speration of individu The measurement grades a performed clustering on three levels of specificity:

1. Correct assignment of individual object to their cluster.
2. Separation of an individual cluster.
3. Overall cohesion and separation of the clustering.

The Silhouette coefficient for an individual object is based on the cohesion within its cluster and the separation from the closest neighboring cluster. A supporting illustration is given in Figure 15. The intra-cluster cohesion for an object i of cluster A is the average dissimilarity of i to all other objects in A, further refered to as $a(i)$. The seperation of object i in cluster A is described as $b(i) = minimum_{C \neq A} d(i, C)$

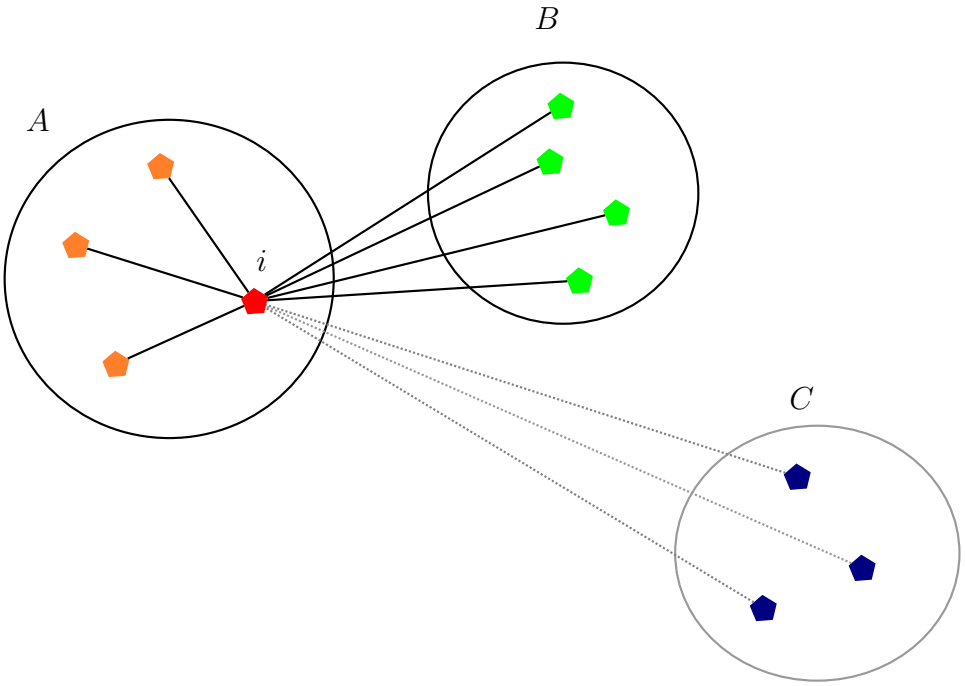


Figure 15: Silhouette coefficient for the object i in cluster A.

where $d(i, C)$ is the average dissimilarity of i to all objects in C. Furthermore let us define cluster B with $minimum_{C \neq A} d(i, C)$ the neighbor cluster of object i in A. The silhouette value for the i-th point, $s(i)$, is then defined as:

$$s(i) = \frac{a(i) - b(i)}{max\{a(i), b(i)\}}$$

The silhouette value ranges from -1 to 1, where 1 indicating a very good cluster assignment for object i and -1 indicates a misclassification of i. A silhouette value around 0 indicates that object i lies in between cluster A and B.

The Silhouette coefficient for an individual cluster C is now calculated as the average Silhouette value of all elements in the cluster; $s(C) = \frac{1}{|C|} \sum s(i)$ with $i_c \in C$.

The Silhouette coefficient for the overall clustering result is the average of the silhouette value of the individual clusters, thus $s(clustering) = \frac{1}{k} \sum s(C)$ with for each C in the clustering.

To identify the optimal number of clusters k in our fanny clustering, an elbow method was used. Here, multiple fanny clustering runs with increasing number of k is plotted against the average silhouette value.

After deciding on the optimal cluster for each of the two matrices, the cluster can be mapped back on the original data. On this resulting matrix, analysis of variance (ANOVA) can be performed. We used here a chi-squared test. Our approach allows to classify an element multiple times, therefore the degree of freedom had to be set to 1.

6.3 Results

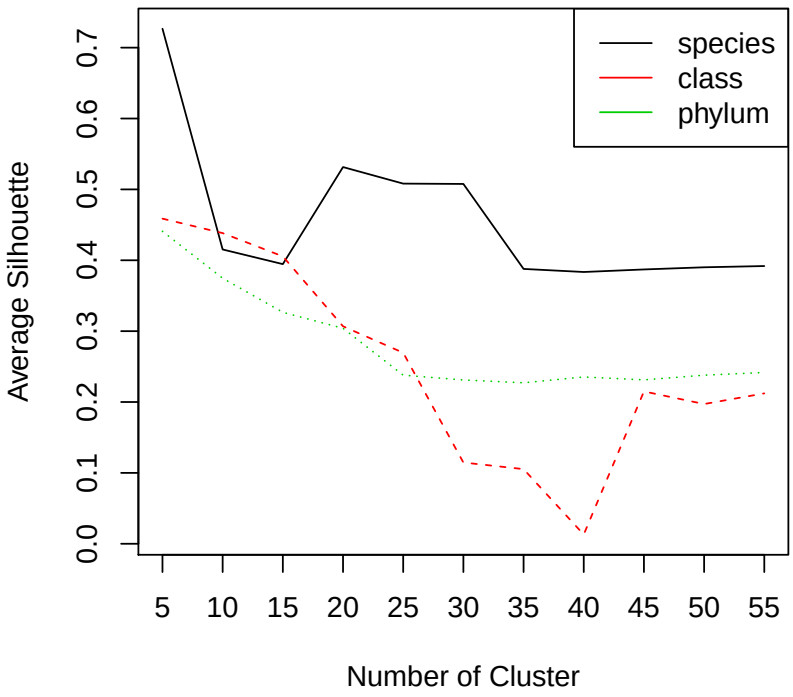


Figure 16: Elbow method for the enzyme PAM clustering. On the x-axis k is given, on the y-axis the average silhouette value is plotted. The three OTU levels, species (black), class (red) and phylum (green) are shown.

We will here only show the results for the OUTO3 sample. Figure 16 shows the elbow function for the enzyme matrix. It ranges from 5 to 55 separate clusters, the average silhouette value is calculated for the multiples of 5. The best silhouette values are best for the smallest amount of clusters. Besides 5 clusters on the species level, none of the silhouette values is above 0.5. A threshold of 0.5 is usually considered to indicate a mediocre structure of the data.

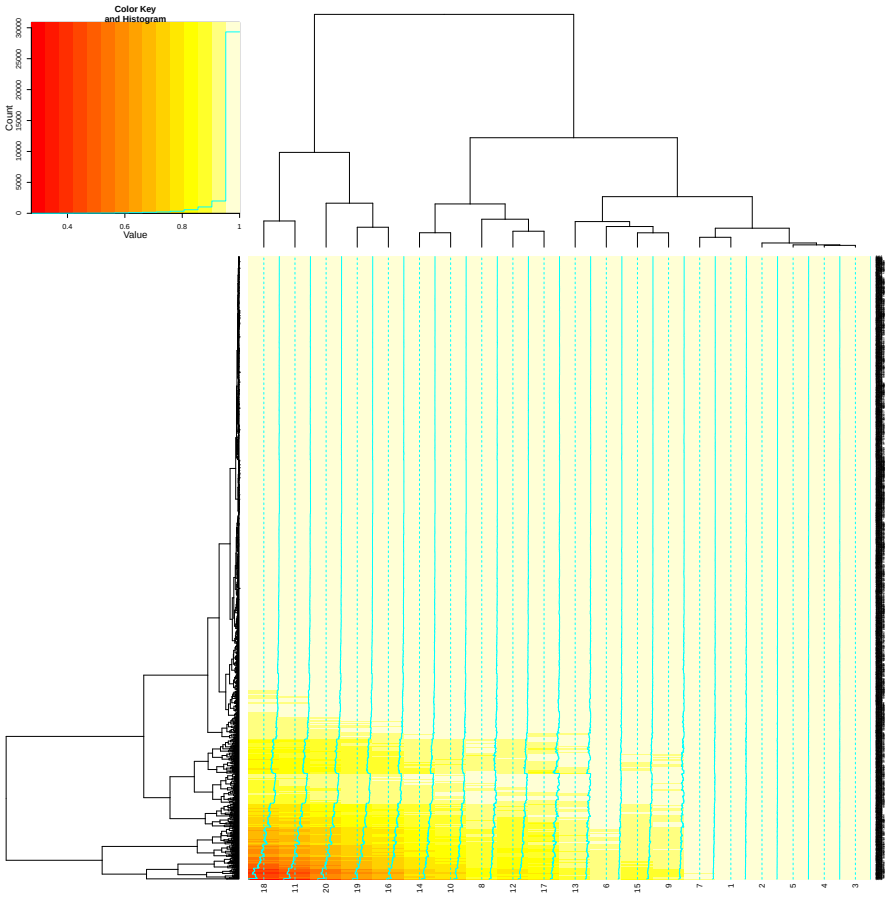


Figure 17: OUTO3, Enzymes on species level, 20 clusters. Negative heatmap of the inner product between the cluster medians and all other enzymes is shown.

A negative heatmap, Figure 17, shows the medians for each cluster and their inner product to all enzymes in the Enzyme $\times$ Enzyme matrix. Many times, the inner product is 0, showing that the two enzymes had not a single OTU in common.

The elbow curves for the OTU's for different taxonomic levels are shown in Figure 18. PAM was applied with clusters between 2 and 10. The overall number of different OTUs is smaller. Thus, a smaller number of clusters is reasonable. Overall, higher average silhouette values are reached, indicating that the clusters capture an underlying structure of enzyme profiles for the taxons.

A negative heatmap, Figure 19, shows the medians for each cluster and their inner product to all OTU in the OTU $\times$ OTU matrix. Although 0 is observed, the frequency is less. Especially, no two cluster medians are essentially the same.

As the clustering apporach has not lead to any significant result for the enzymes, a

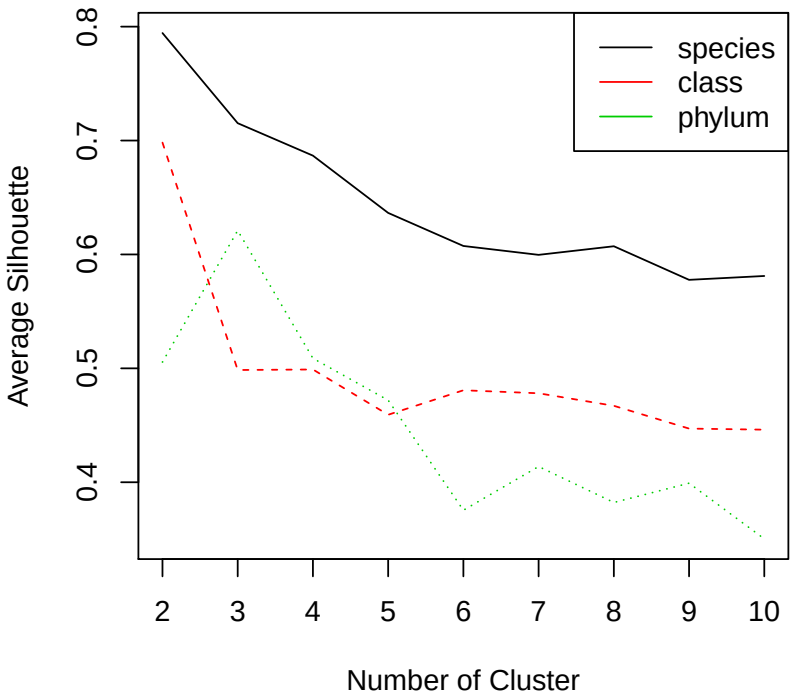


Figure 18: Elbow method for the OTU's PAM clustering. On the x-axis k is given, on the y-axis the average silhouette value is plotted. The three OTU levels, species (black), class (red) and phylum (green) are shown.

back mapping and further statistical test did not lead to any further results.

6.4 Discussion

Here we presented an outline and first results for a possible taxonomic distribution of pathways detection method.

The clustering of the Enzyme $\times$ Enzyme matrix has not lead to strong silhouette values, indicating no structure underlying this data. An increase of the number of clusters might reveal specific, small groups of similar enzymes. However, most of the pairwise comparisons of the enzymes showed that they do not have any OTUs in common. Different kernel methods could also be applied to the initial data. The dot product, however, already is a non-symmetric measurement. The matrix data is too sparse for this novel strategy. A different annotation approach or more sequencing

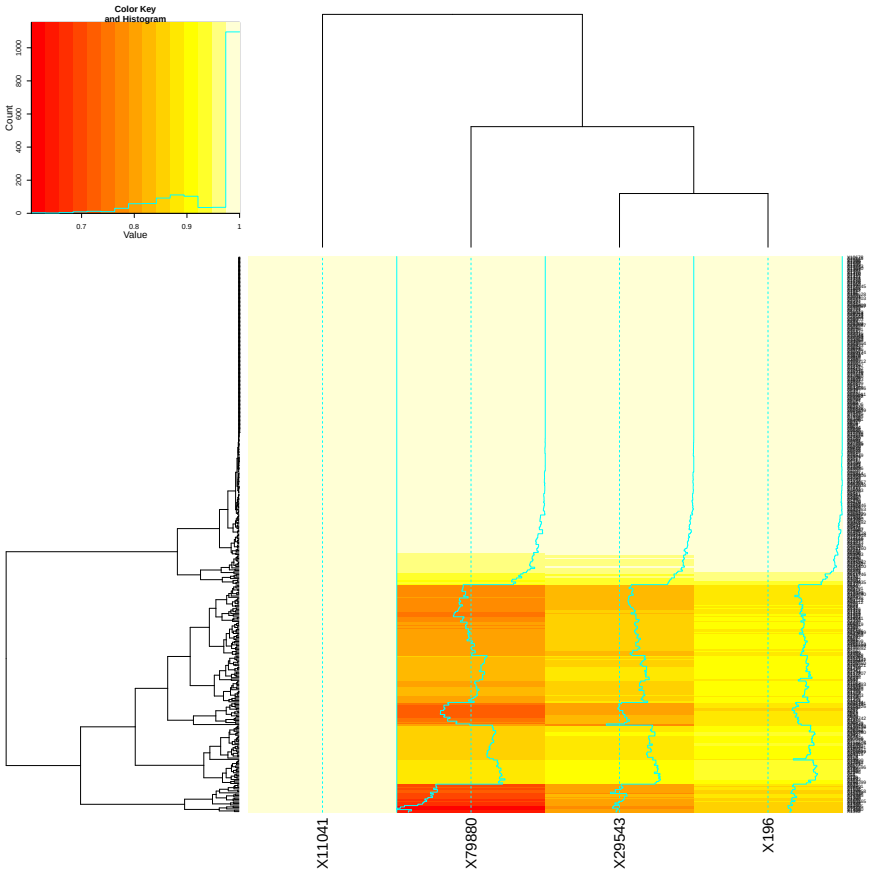


Figure 19: OUTO3, OTUs on species level, 4 clusters. Negative heatmap of the inner product between the cluster medians and all other enzymes is shown.

data might be a solution.

The clustering of the OTU × OTU matrix is promising. Clusters separate well and medians are reasonable.

For testing purposes, simulated data could be generated and a full analysis performed. Further validation of this strategy could be done by using the leave-one-out cross-validation strategy. Furthermore, the chi-square test should be replaced. The chi-squared test is an approximation of the Fisher's exact test. The Fisher's exact test can, nowadays, be computed within reasonable time, making the chi-square test somewhat obsolete [PB10]. Moreover, different clustering techniques, such as the overlapping clustering of fanny [KR90] could relax constraints that are not supported by biological environments. Because, for example an organism is most likely involved in multiple metabolic processes.

7 Concluding Remarks

We here characterized novel metagenomic samples from the relatively unexplored deep biosphere. We applied several techniques to analyze the taxonomic composition and metabolic capacity of the microbial communities.

In Chapter 3 we presented several basic and more advanced techniques to assessing the adequacy of the sample coverage and annotation in the novel metagenomic samples. We were not able to use indexes, such as the Shannon index, although they are commonly used in biodiversity analysis. These indexes are supposed to be used with single copy genes. Our metagenomic data does not fulfill this criterion. However, we were able to modify the rarefaction curve to our needs. Furthermore, we extended its application field by also using it for estimating the functional annotations. We have to keep in mind though, that, as opposed to the original rarefaction curve, we now measure two factors simultaneously. The original function of rarefaction curve assessed whether or not the sample size was adequate for the habitat. The use of the rarefaction curve as presented here is also greatly affected by the annotation method. With a different reference database, or another, not homology-based annotation method, other results might have occurred. Methods like MetaID [SG13], an alignment-free n-gram approach designed for short reads, could be considered in the future.

For the here shown novel metagenomic samples of deep biosphere the sample size should at least covers the genomic material to an extend that an assembly is possible in the future. Otherwise biases due to incorrect annotations arise.

In Chapter 5 we provided and extensive visualization of the metabolic capacities in our metagenomic samples. We are currently collaborating with domain experts in order to investigate further novel interpretations of the metabolic pathways.

Further in Chapter 4 we conducted a comparative analysis of the deep biosphere metagenomes and publicly available metagenomes as well as individual genomes. We using two variants of PCA, the standard PCA and a variation, the sparse PCA. The use of sparse PCA conducted results similar to PCA, however with the advert of easier to interpret PCs.

We were not able to confirm initial assumptions of Fungi or Archaea metabolism in the deep bedrock biosphere.

Finally, in Chapter 6 we outline a novel approach to learn the taxonomic distribution of pathways. First results are promising and further improvement on the approach

as well as the initial annotations process could lead to identifying pathways implemented by specific taxonomic units.

The here presented work has given preliminary understanding of the deep bedrock biosphere. Further analysis and further sampling would be necessary to establish a profound knowledge in order to assess risks connected to altering the environment.

As a last point, it should be noted that further results are available at:
http://www.cs.helsinki.fi/group/urenzyme/deepfun/

List of Figures

List of Tables

References

ABF+03 Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., Uddin, A. and Zygouri, C., Prints and its automatic supplement, preprints. *Nucleic Acids Research*, 31,1(2003), pages 400–402. URL `http://nar.oxfordjournals.org/content/31/1/400`.

AGM+90 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. and Lipman, D. J., Basic local alignment search tool. *J. Mol. Biol.*, 215,3(1990), pages 403–410. URL `http://dx.doi.org/10.1016/S0022-2836(05)80360-2`.

AJL+07 Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. and Walter, P., *Molecular Biology of the Cell*. Garland Science, fifth edition, 2007.

AKB09 Allen, B., Kon, M. and Bar, Y., A new phylogenetic diversity measure generalizing the shannon index and its application to phyllostomid bats. *The American Naturalist*, 174,2(2009), pages 236–243. URL `http://dx.doi.org/10.1086/600101`.

ALS95 Amann, R., Ludwig, W. and Schleifer, K., Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol. Rev.*, 59,1(1995), pages 143–69.

ATK+13 Altman, T., Travers, M., Kothari, A., Caspi, R. and Karp, P., A systematic comparison of the metacyc and kegg pathway databases.

BMC Bioinformatics, 14,1(2013), page 112. URL http://www.biomedcentral.com/1471-2105/14/112.

AXL+12 Arndt, D., Xia, J., Liu, Y., Zhou, Y., Guo, A. C., Cruz, J. A., Sinelnikov, I., Budwill, K., Nesbø, C. L. and Wishart, D. S., Metagenassist: a comprehensive web server for comparative metagenomics. *Nucleic Acids Research*, 40,W1(2012), pages W88–W95. URL http://nar.oxfordjournals.org/content/40/W1/W88.

BAK+00 Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., Spudich, E. N. and DeLong, E. F., Bacterial rhodopsin: Evidence for a new type of phototrophy in the sea. *Science*, 289,5486(2000), pages 1902–1906. URL http://www.sciencemag.org/content/289/5486/1902.

BCA11 Bates, J. T., Chivian, D. and Arkin, A. P., Glamm: Genome-linked application for metabolic maps. *Nucleic Acids Res*, 39,Suppl. 2(2011). URL http://nar.oxfordjournals.org/content/early/2011/05/28/nar.gkr433.full.

Ber96 Berg, R. D., The indigenous gastrointestinal microflora. *Trends in Microbiology*, 4,11(1996), pages 430–435. URL http://dx.doi.org/10.1016/0966-842X(96)10057-3.

BH95 Benjamini, Y. and Hochberg., Y., Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B (Methodological)*, 57,1(1995), pages 289–300.

Bis06 Bishop, C. M., *Pattern Recognition and Machine Learning (Information Science and Statistics)*. Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.

BKS+06 Berven, F. S., Karlsen, O. A., Straume, A. H., Flikka, K., Murrell, J. C., Fjellbirkeland, A., Lillehaug, J. R., Eidhammer, I. and Jensen, H. B., Analysing the outer membrane subproteome of methylococcus capsulatus (bath) using proteomics and novel biocomputing tools. *Arch Microbiol*, 184,6(2006), pages 362–377. URL http://dx.doi.org/10.1007/s00203-005-0055-7.

BLH06 Bagos, P., Liakopoulos, T. and Hamodrakas, S., Algorithms for incorporating prior topological information in hmms: application to transmembrane proteins. *BMC Bioinformatics*, 7,1(2006), page 189. URL http://dx.doi.org/10.1186/1471-2105-7-189.

Bri14 Britannica, E., Venn diagramm entry - encyclopedia britannica., 2014. URL http://global.britannica.com/EBchecked/topic/625448/Venn-diagram.

BS09 Brady, A. and Salzberg, S. L., Phymm and phymmbl: metagenomic phylogenetic classification with interpolated markov models. *Nature Methods*, 6,9(2009), pages 673–676. URL http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=43950542&site=ehost-live&scope=site.

BvdBC+00 Baker, W., van den Broek, A., Camon, E., Hingamp, P., Sterk, P., Stoesser, G. and Tuli, M. A., The embl nucleotide sequence database. *Nucleic Acids Research*, 28,1(2000), pages 19–23. URL http://nar.oxfordjournals.org/content/28/1/19.

CAB+14 Caspi, R., Altman, T., Billington, R., Dreher, K., Foerster, H., Fulcher, C. A., Holland, T. A., Keseler, I. M., Kothari, A., Kubo, A., Krummenacker, M., Latendresse, M., Mueller, L. A., Ong, Q., Paley, S., Subhraveti, P., Weaver, D. S., Weerasinghe, D., Zhang, P. and Karp, P. D., The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 42,D1(2014), pages D459–D471. URL http://nar.oxfordjournals.org/content/42/D1/D459.

CF11 Campbell, M. K. and Farrell., S. O., *Biochemistry*. Brooks Cole, 2011.

CZC+13 Chen, W., Zhang, C. K., Cheng, Y., Zhang, S. and Zhao, H., A comparison of methods for clustering 16s rrna sequences into otus. *PLoS ONE*, 8,8(2013). URL http://dx.doi.org/10.1371/journal.pone.0070837.

dEJL07 d'Aspremont, A., El Ghaoui, L., Jordan, M. and Lanckriet, G., A direct formulation of sparse PCA using semidefinite programming. *SIAM Review*, 49,3(2007).

Dij59 Dijkstra, E. W., A note on two problems in connexion with graphs. *Numerische Mathematik 1*, 1,1(1959), pages 269–271. URL http://www.cs.yale.edu/homes/lans/readings/routing/dijkstra-routing-1959.pdf.

DJL+14 Darling, A. E., Jospin, G., Lowe, E., Matsen, IV, F. A., Bik, H. M. and Eisen, J. A., Phylosift: phylogenetic analysis of genomes and metagenomes. *PeerJ*, 2,e(2014), page e243. URL http://dx.doi.org/10.7717/peerj.243.

DKG+09 Diaz, N., Krause, L., Goesmann, A., Niehaus, K. and Nattkemper, T., Tacoa - taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, 10,1(2009), page 56. URL http://www.biomedcentral.com/1471-2105/10/56.

DSFM+12 Dos Santos, P., Fang, Z., Mason, S., Setubal, J. and Dixon, R., Distribution of nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC Genomics*, 13,1(2012), page 162. URL http://www.biomedcentral.com/1471-2164/13/162.

DSHB98 Dandekar, T., Snel, B., Huynen, M. and Bork, P., Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in Biochemical Sciences*, 23,9(1998), pages 324–328. URL http://dx.doi.org/10.1016/S0968-0004(98)01274-2.

DW09 Doolittle WF, Z. O., On the origin of prokaryotic species. *Genome Research*, 19,5(2009), pages 744–56.

Efr03 Efron, B., Second thoughts on the bootstrap. *Statistical Science*, 18,2(2003), pages 135–140. URL http://projecteuclid.org/euclid.ss/1063994968.

EMB14 EMBL–EBI, Biomodels database, bacillus cereus., 2014. URL http://www.ebi.ac.uk/biomodels-main/BMID000000060972.

GHSM12 Garmendia, L., Hernandez, A., Sanchez, M. B. and Martinez, J. L., Metagenomics and antibiotics. *Clinical Microbiology and Infection*, 18,Suppl. 4(2012), pages 27–31. URL http://dx.doi.org/10.1111/j.1469-0691.2012.03868.x.

GWW+10 Glass, E., Wilkening, J., Wilke, A. et al., Using the metagenomics rast server (mg-rast) for analyzing shotgun metagenomes. *Cold Spring Harb Protoc*, 2010,1(2010).

HDG+10 Hemme, C. L., Deng, Y., Gentry, T. J., Fields, M. W., Wu, L., Barua, S., Barry, K. and other, Metagenomic insights into evolution of a heavy metal-contaminated groundwater microbial community. *The ISME journal*, 4,5(2010), pages 660–672.

HE13 Headd, B. and Engel, A. S., Evidence for niche partitioning revealed by the distribution of sulfur oxidation genes collected from areas of a terrestrial sulfidic spring with differing geochemical conditions. 79,4(2013), pages 1171–1182.

Her12 Herrmann., Y. N., Comparison of metagenomic data by identifying enriched pathways in metabolic networks. M.Sc. thesis, Universitat Bielefeld, 2012.

HH92 Henikoff, S. and Henikoff, J. G., Amino acid substitution matrices from protein blocks. *Proceedings of the National Academy of Sciences*, 89,22(1992), pages 10915–10919. URL `http://www.pnas.org/content/89/22/10915`.

HHR10 Hawkins, R. D., Hon, G. C. and Ren, B., Next-generation genomics: an integrative approach. *Nat Rev Genet*, 11,7(2010), pages 476–486. URL `http://dx.doi.org/10.1038/nrg2795`.

HMW+11 Huson, D. H., Mitra, S., Weber, N., Ruscheweyh, H.-J. and Schuster, S. C., Integrative analysis of environmental sequences using megan4. *Genome Research*, 21,9(2011), pages 1552–1560.

HRB+98 Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. and Goodman, R. M., Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5,10(1998), pages R245–R249. URL `http://dx.doi.org/10.1016/S1074-5521(98)90108-9`.

Ill14 Illumina, Illumina product information, 2014. URL `http://www.illumina.com/systems/sequencing.ilmn`.

Ins14 Institute., D. J. G., Marine anammox bioreactor enriched for scalindua species., 2014. URL http://genome.jgi.doe.gov/SedMa_2017108002/SedMa_2017108002.info.html.

JBD+09 Jain, E., Bairoch, A., Duvaud, S., Phan, I., Redaschi, N., Suzek, B. E., Martin, M. J. and other, Infrastructure for the life sciences: design and implementation of the uniprot website. *BMC Bioinformatics*, 10,1(2009), page 136.

JKS06 Junker, B., Klukas, C. and Schreiber, F., Vanted: A system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7,1(2006), page 109. URL http://www.biomedcentral.com/1471-2105/7/109.

Jol86 Jolliffe, I., *Principal Component Analysis*. Springer Verlag, 1986.

KAG+08 Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T. and Yamanishi, Y., Kegg for linking genomes to life and the environment. *Nucleic Acids Research*, 36,Suppl 1(2008), pages D480–D484. URL http://nar.oxfordjournals.org/content/36/suppl_1/D480.

Kan96 Kanehisa, M., Towards pathway engineering: a new database of genetic and molecular pathways. *Science & Technology Japan*, 1996,59(1996), pages 34–38. URL http://www.kanehisa.jp/docs/archive/stj.pdf.

KCL+08 Kunin, V., Copeland, A., Lapidus, A., Mavromatis, K. and Hugenholtz, P., A bioinformatician's guide to metagenomics. *Microbiology and Molecular Biology Reviews*, 72,4(2008), pages 557–578. URL http://mmbr.asm.org/content/72/4/557.

KGS+14 Kanehisa, M., Goto, S., Sato, Y., Kawashima, M., Furumichi, M. and Tanabe, M., Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Acids Research*, 42,D1(2014), pages D199–D205. URL http://nar.oxfordjournals.org/content/42/D1/D199.abstract.

KR90 Kaufman, L. and Rousseeuw, P. J., *Finding groups in data: an introduction to cluster analysis*, volume 1. John Wiley and Sons, New York, 1990.

KSL+13 Koboldt, D. C., Steinberg, K. M., Larson, D. E., Wilson, R. K. and Mardis, E. R., The next-generation sequencing revolution and its impact on genomics. *Cell*, 155,1(2013), pages 27–38. URL http://dx.doi.org/10.1016/j.cell.2013.09.006.

Kut11 Kutschera, U., From the scala naturae to the symbiogenetic and dynamic tree of life. *Biology Direct*, 6,1(2011), page 33. URL http://www.biology-direct.com/content/6/1/33.

Lev11 Levering, J., *Systems biology of the central metabolism of Streptococcus pyogenes*. Ph.D. thesis, University of Heidelberg, Germany, 2011.

LYY+11 Leung, H. C. M., Yiu, S. M., Yang, B., Peng, Y., Wang, Y., Liu, Z., Chen, J., Qin, J., Li, R. and Chin, F. Y. L., A robust and accurate binning algorithm for metagenomic sequences with arbitrary species abundance ratio. *Bioinformatics*, 27,11(2011), pages 1489–1495. URL http://bioinformatics.oxfordjournals.org/content/27/11/1489.

MCC+12 Markowitz, V., Chen, I., Chu, K. et al., Img/m: the integrated metagenome data management and comparative analysis system. *Nucleic Acids Res.*, 40,D123–9(2012).

MCP+12 Markowitz, V. M., Chen, I.-M. A., Palaniappan, K., Chu, K., Szeto, E., Grechkin, Y., Ratner, A., Jacob, B., Huang, J., Williams, P., Huntemann, M., Anderson, I., Mavromatis, K., Ivanova, N. N. and Kyrpides, N. C., Img: the integrated microbial genomes database and comparative analysis system. *Nucleic Acids Research*, 40,D1(2012), pages D115–D122. URL http://nar.oxfordjournals.org/content/40/D1/D115.

MCRFAS12 Montoya, L., Celis, L., Razo-Flores, E. and Alpuche-Solís, Á., Distribution of co2 fixation and acetate mineralization pathways in microorganisms from extremophilic anaerobic biotopes. *Extremophiles*, 16,6(2012), pages 805–817. URL http://dx.doi.org/10.1007/s00792-012-0487-3.

ME13 Matsen, F. A. and Evans, S. N., Edge principal components and squash clustering: Using the special structure of phylogenetic placement data for sample comparison. *PLoS ONE*, 8,3(2013). URL http://dx.doi.org/10.1371\%2Fjournal.pone.0056859.

MIB+07 Mavromatis, K., Ivanova, N., Barry, K., Shapiro, H., Goltsman, E., McHardy, A. C., Rigoutsos, I., Salamov, A., Korzeniewski, F., Land, M., Lapidus, A., Grigoriev, I., Richardson, P., Hugenholtz, P. and Kyrpides, N. C., Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Meth*, 4,6(2007), pages 495–500. URL `http://dx.doi.org/10.1038/nmeth1043`.

MIO+07 Moriya, Y., Itoh, M., Okuda, S., Yoshizawa, A. C. and Kanehisa, M., Kaas: an automatic genome annotation and pathway reconstruction server. *Nucleic Acids Research*, 35,suppl 2(2007), pages W182–W185. URL `http://nar.oxfordjournals.org/content/35/suppl_2/W182.abstract`.

MMG12 Mande, S. S., Mohammed, M. H. and Ghosh, T. S., Classification of metagenomic sequences: methods and challenges. *Briefings in Bioinformatics*, 13,6(2012), pages 669–681. URL `http://bib.oxfordjournals.org/content/13/6/669`.

MMT+07 McHardy, A. C., Martín, H. G., Tsirigos, A., Hugenholtz, P. and Rigoutsos, I., Accurate phylogenetic classification of variable-length dna fragments. *Nature Methods*, 4,1(2007), pages 63–72.

MvRTB12 Medema, M., van Raaphorst, R., Takano, E. and Breitling, R., Computational tools for the synthetic design of biochemical pathways. *Nat Rev Microbiol*, 10,3(2012), pages 191–202.

NHTS12 Namiki, T., Hachiya, T., Tanaka, H. and Sakakibara, Y., Metavelvet: an extension of velvet assembler to de novo metagenome assembly from short sequence reads. *Nucleic Acids Research*, 40,20(2012). URL `http://nar.oxfordjournals.org/content/early/2012/07/19/nar.gks678`.

PB10 Parks, D. H. and Beiko, R. G., Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26,6(2010), pages 715–721. URL `http://bioinformatics.oxfordjournals.org/content/26/6/715`.

PBGLVC+13 Perez-Brocal, V., Garcia-Lopez, R., Vazquez-Castellanos, J. F., Nos, P., Beltran, B., Latorre, A. and Moya, A., Study of the viral and

microbial communities associated with crohn/'s disease: A metagenomic approach. *Clin Trans Gastroenterol*, 4,6(2013). URL `http://dx.doi.org/10.1038/ctg.2013.9`.

Pea01 Pearson, K., On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2,1(1901), pages 559–572.

PHP+11 Patil, K. R., Haider, P., Pope, P. B., Turnbaugh, P. J., Morrison, M., Scheffer, T. and McHardy, A. C., Taxonomic metagenome sequence assignment with structured output models. *Nat Methods*, 8,3(2011).

PLYC11 Peng, Y., Leung, H. C. M., Yiu, S. M. and Chin, F. Y. L., Meta-idba: a de novo assembler for metagenomic data. *Bioinformatics*, 27,13(2011), pages i94–i101. URL `http://bioinformatics.oxfordjournals.org/content/27/13/i94`.

PS10 Paszkiewicz, K. and Studholme, D. J., De novo assembly of short sequence reads. *Briefings in Bioinformatics*, 11,5(2010), pages 457–472. URL `http://bib.oxfordjournals.org/content/11/5/457`.

PT12 Prakash, T. and Taylor, T. D., Functional assignment of metagenomic data: challenges and applications. *Briefings in Bioinformatics*, 13,6(2012), pages 711–727.

QPY+13 Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. and Glöckner, F. O., The silva ribosomal rna gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, 41,D1(2013), pages D590–D596. URL `http://nar.oxfordjournals.org/content/41/D1/D590`.

RHS+07 Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M. and Venter, J. C., The sorcerer ii global

ocean sampling expedition: Northwest atlantic through eastern tropical pacific. *PLoS Biol*, 5,3(2007). URL `http://dx.doi.org/10.1371/journal.pbio.0050077`.

RJH+12 Rohn, H., Junker, A., Hartmann, A., Grafahrend-Belau, E., Treutler, H., Klapperstuck, M., Czauderna, T., Klukas, C. and Schreiber, F., Vanted v2: a framework for systems biology applications. *BMC Systems Biology*, 6,1(2012), page 139. URL `http://www.biomedcentral.com/1752-0509/6/139`.

Rou87 Rousseeuw, P. J., Silhouette: a graphical aid to interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20,0(1987), pages 53–65.

RUC+10 Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V. and Jackson, R. B., *Campbell Biology (9th Edition)*. Benjamin Cummings, 9th edition, 2010.

RUC+11 Reece, J. B., Urry, L. A., Cain, M. L., Wasserman, S. A., Minorsky, P. V. and Jackson, R. B., *Campbell Biology*. Benjamin Cummings / Pearson, 9th edition, 2011.

Sav77 Savage, D. C., Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology*, 31,1(1977), pages 107–133. URL `http://dx.doi.org/10.1146/annurev.mi.31.100177.000543`.

SCdC+10 Sigrist, C. J. A., Cerutti, L., de Castro, E., Langendijk-Genevaux, P. S., Bulliard, V., Bairoch, A. and Hulo, N., Prosite, a protein domain database for functional characterization and annotation. *Nucleic Acids Research*, 38,Suppl 1(2010), pages D161–D166. URL `http://nar.oxfordjournals.org/content/38/suppl_1/D161`.

SCL+11 Sun, S., Chen, J., Li, W. et al., Community cyberinfrastructure for advanced microbial ecology research and analysis: the camera resource. *Nucleic Acids Res.*, 39,Suppl. 1(2011), pages D546–51.

SD12 Simpson, J. T. and Durbin, R., Efficient de novo assembly of large genomes using compressed data structures. *Genome Research*, 22,3(2012), pages 549–556. URL `http://genome.cshlp.org/content/22/3/549`.

SE11 Schmieder, R. and Edwards, R., Fast identification and removal of sequence contamination from genomic and metagenomic datasets. *PloS one*, 6,3(2011). URL http://europepmc.org/abstract/MED/21408061.

SG13 Srinivasan, S. and Guda, C., Metaid: A novel method for identification and quantification of metagenomic samples. *BMC Genomics*, 14,Suppl 8(2013), page S4. URL http://dx.doi.org/10.1186/1471-2164-14-S8-S4.

SH05 Schloss, P. and Handelsman, J., Metagenomics for studying unculturable microorganisms: cutting the gordian knot. *Genome Biology*, 6,8(2005), page 229.

SH08 Schloss, P. and Handelsman, J., A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics*, 9,1(2008), page 34. URL http://www.biomedcentral.com/1471-2105/9/34.

SK85 Staley, J. and Konopka, A., Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol*, 39,1(1985), pages 321–346.

SLD+12 Sangwan, N., Lata, P., Dwivedi, V., Singh, A., Niharika, N., Kaur, J., Anand, S., Malhotra, J., Jindal, S., Nigam, A., Lal, D., Dua, A., Saxena, A., Garg, N., Verma, M., Kaur, J., Mukherjee, U., Gilbert, J. A., Dowd, S. E., Raman, R., Khurana, P., Khurana, J. P. and Lal, R., Comparative metagenomic analysis of soil microbial communities across three hexachlorocyclohexane contamination levels. *PLoS ONE*, 7,9(2012). URL http://dx.doi.org/10.1371/journal.pone.0046219.

ST03 Storey, J. and Tibshirani, R., Statistical significance for genomewide studies. *Proceedings of the National Academy of Sciences of the United States of America*, 100,16(2003), pages 9440–9445. URL http://europepmc.org/abstract/MED/12883005.

STDY10 Shah, N., Tang, H., Doak, T. G. and Ye, Y. *Comparing bacterial communities inferred from 16s rRNA gene sequencing and shot-*

gun metagenomics, pages 165–176. World scientific, 2010. URL http://dx.doi.org/10.1142/9789814335058%7B%5C_%7D0018.

TCH+04 Tyson, G., Chapman, J., Hugenholtz, P. et al., Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428,6978(2004), pages 37–43.

TG12 Temperton, B. and Giovannoni, S. J., Metagenomics: microbial diversity through a scratched lens. *Current Opinion in Microbiology*, 15,5(2012), pages 605–612. URL http://dx.doi.org/10.1016/j.mib.2012.07.001.

TWL+04 Teeling, H., Waldmann, J., Lombardot, T., Bauer, M. and Glockner, F., Tetra: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in dna sequences. *BMC Bioinformatics*, 5,1(2004), page 163. URL http://www.biomedcentral.com/1471-2105/5/163.

UIL+13 Uroz, S., Ioannidis, P., Lengelle, J., Cébron, A., Morin, E., Buée, M. and Martin, F., Functional assays and metagenomic analyses reveals differences between the microbial communities inhabiting the soil horizons of a norway spruce plantation. *PLoS ONE*, 8,2(2013), page e55929. URL http://dx.doi.org/10.1371/journal.pone.0055929.

vLS59 von Linné, C. and Salvius, L., *Systema naturae per regna tria naturae :secundum classes, ordines, genera, species, cum characteribus, differentiis, synonymis, locis.*, volume 1. Holmiae Impensis Direct. Laurentii Salvii, 1759. URL http://www.biodiversitylibrary.org/item/10278.

VRH+04 Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H. and Smith, H. O., Environmental genome shotgun sequencing of the sargasso sea. *Science*, 304,5667(2004), pages 66–74. URL http://www.sciencemag.org/content/304/5667/66.

WCW98 Whitman, W. B., Coleman, D. C. and Wiebe, W. J., Prokaryotes: the

unseen majority. *Proc Natl Acad Sci U S A*, 95,12(1998), pages 6578–6583.

WGF10 Wooley, J. C., Godzik, A. and Friedberg, I., A primer on metagenomics. *PLoS Comput Biol*, 6,2(2010). URL `http://dx.doi.org/10.1371`.

WPT+11 Waldron, L., Pintilie, M., Tsao, M.-S., Shepherd, F. A., Huttenhower, C. and Jurisica, I., Optimized application of penalized regression methods to diverse genomic data. *Bioinformatics*, 27,24(2011), pages 3399–3406.

YLO+11 Yamada, T., Letunic, I., Okuda, S., Kanehisa, M. and Bork, P., ipath2.0: interactive pathway explorer. *Nucleic Acids Res.*, 26,Suppl. 2(2011). URL `http://www.ncbi.nlm.nih.gov/pubmed/21546551`.

YNR+10 Yooseph, S., Nealson, K., Rusch, D., McCrow, J., Dupont, C., Kim, M., Johnson, J., Montgomery, R., Ferriera, S., Beeson, K. et al., Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, 468,7320(2010), pages 60–66.

ZHT04 Zou, H., Hastie, T. and Tibshirani, R., Sparse principal component analysis. *Journal of Computational and Graphical Statistics*, 15,2(2004), pages 265–286. URL `http://www.stanford.edu/~hastie/Papers/spc_jcgs.pdf`.

ZSK99 Zweier, J. L., Samouilov, A. and Kuppusamy, P., Non-enzymatic nitric oxide synthesis in biological systems. *Biochimica et Biophysica Acta (BBA) - Bioenergetics*, 1411,2–3(1999), pages 250–262. URL `http://dx.doi.org/10.1016/S0005-2728(99)00018-3`.